

This Page Is Inserted by IFW Operations
and is not a part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

**As rescanning documents *will not* correct images,
please do not report the images to the
Image Problem Mailbox.**

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

<p>(51) International Patent Classification⁶ : C12Q 1/68</p>	<p>A1</p>	<p>(11) International Publication Number: WO 95/21944 (43) International Publication Date: 17 August 1995 (17.08.95)</p>
<p>(21) International Application Number: PCT/US95/01863 (22) International Filing Date: 14 February 1995 (14.02.95) (30) Priority Data: 08/195,485 14 February 1994 (14.02.94) US (60) Parent Application or Grant (63) Related by Continuation US 08/195,485 (CIP) Filed on 14 February 1994 (14.02.94) (71) Applicant (for all designated States except US): SMITHKLINE BEECHAM CORPORATION [US/US]; Corporate Intellectual Property, UW2220, 709 Swedeland Road, P.O. Box 1539, King of Prussia, PA 19406-0939 (US). (72) Inventors; and (75) Inventors/Applicants (for US only): ROSENBERG, Martin [US/US]; 241 Mingo Road, Royersford, PA 19468 (US). DEBOUCK, Christine [BE/US]; 667 Pugh Road, Wayne, PA 19087 (US). BERGSMA, Derk [US/US]; 271 Irish Road, Berwyn, PA 19312 (US).</p>		<p>(74) Agents: JERVIS, Herbert, H. et al.; SmithKline Beecham Corporation, Corporate Intellectual Property, UW2220, 709 Swedeland Road, P.O. Box 1539, King of Prussia, PA 19406-0939 (US). (81) Designated States: JP, US, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE). Published With international search report.</p>
<p>(54) Title: DIFFERENTIALLY EXPRESSED GENES IN HEALTHY AND DISEASED SUBJECTS (57) Abstract The present invention involves methods and compositions for identifying genes which are differentially expressed in a normal healthy animal and an animal having a selected disease or infection, and methods for diagnosing diseases or infections characterized by the presence of those genes, despite the absence of knowledge about the gene or its function. The methods involve the use of a composition suitable for use in hybridization which consists of a solid surface on which is immobilized at pre-defined regions thereon a plurality of defined oligonucleotide/polynucleotide sequences for hybridization. Each sequence comprises a fragment of an EST isolated from an identified DNA library prepared from tissue or cell samples of a healthy animal, an animal with a selected disease or infection, and any combination thereof. Differences in hybridization patterns produced through use of this composition and the specified methods enable diagnosis of disease based on differential expression of genes of unknown function, and enable the identification of those genes and the proteins encoded thereby.</p>		

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	GB	United Kingdom	MR	Mauritania
AU	Australia	GE	Georgia	MW	Malawi
BB	Barbados	GN	Guinea	NE	Niger
BE	Belgium	GR	Greece	NL	Netherlands
BF	Burkina Faso	HU	Hungary	NO	Norway
BG	Bulgaria	IE	Ireland	NZ	New Zealand
BJ	Benin	IT	Italy	PL	Poland
BR	Brazil	JP	Japan	PT	Portugal
BY	Belarus	KE	Kenya	RO	Romania
CA	Canada	KG	Kyrgyzstan	RU	Russian Federation
CF	Central African Republic	KP	Democratic People's Republic of Korea	SD	Sudan
CG	Congo	KR	Republic of Korea	SE	Sweden
CH	Switzerland	KZ	Kazakhstan	SI	Slovenia
CI	Côte d'Ivoire	LI	Liechtenstein	SK	Slovakia
CM	Cameroon	LK	Sri Lanka	SN	Senegal
CN	China	LU	Luxembourg	TD	Chad
CS	Czechoslovakia	LV	Latvia	TG	Togo
CZ	Czech Republic	MC	Monaco	TJ	Tajikistan
DE	Germany	MD	Republic of Moldova	TT	Trinidad and Tobago
DK	Denmark	MG	Madagascar	UA	Ukraine
ES	Spain	ML	Mali	US	United States of America
FI	Finland	MN	Mongolia	UZ	Uzbekistan
FR	France			VN	Viet Nam
GA	Gabon				

differentially expressed genes in healthy and diseased subjects

Cross Reference to Related Applications:

5 This application is a continuation-in-part application of U.S. Serial No. 08/195,485 filed February 14, 1994, the contents of which are incorporated herein by reference.

Field of the Invention

10 The present invention relates to the use of immobilized oligonucleotide/polynucleotide or polynucleotide sequences for the identification, sequencing and characterization of genes which are implicated in disease, infection, or development and the use of such identified genes and the proteins encoded thereby in diagnosis, prognosis, therapy and drug discovery.

15 Background of the Invention

Identification, sequencing and characterization of genes, especially human genes, is a major goal of modern scientific research. By identifying genes, determining their sequences and characterizing their biological function, it is possible
20 to employ recombinant DNA technology to produce large quantities of valuable "gene products", e.g., proteins and peptides. Additionally, knowledge of gene sequences can provide a key to diagnosis, prognosis and treatment of a variety of disease states in plants and animals which are characterized by inappropriate expression and/or repression of selected gene(s) or by the influence of external factors, e.g., carcinogens or teratogens, on gene function. The term disease-associated genes(s) is used herein
25 in its broadest sense to mean not only genes associated with classical inherited diseases, but also those associated with genetic predisposition to disease as well as infectious or pathogenic states resulting from gene expression by infectious agents or the effect on host cell gene expression by the presence of such a pathogen or its products. Locating disease-associated genes will permit the development of
30 diagnostic and prognostic reagents and methods, as well as possible therapeutic regimens, and the discovery of new drugs for treating or preventing the occurrence of such diseases.

Methods have been described for the identification of certain novel
35 gene sequences, referred to as Expressed Sequence Tags (EST) [see, e.g., Adams et al, Science, 252:1651-1656 (1991); and International Patent Application No. WO93/00353, published January 7, 1993]. Conventionally, an EST is a specific cDNA polynucleotide sequence, or tag, about 150 to 400 nucleotides in length, derived from

a messenger RNA molecule by reverse transcription, which is a marker for, and component of, a human gene actually transcribed *in vivo*. However, as used herein an EST also refers to a genomic DNA fragment derived from an organism, such as a microorganism, the DNA of which lacks intron regions.

5 A variety of techniques have been described for identifying particular gene sequences on the basis of their gene products. For example, several techniques are described in the art [see, e.g., International Patent Application No. WO91/07087, published May 30, 1991]. Additionally, known methods exist for the amplification of desired sequences [see, e.g., International Patent Application No. WO91/17271,
10 published November 14, 1991, among others].

 However, at present, there exist no established methods for filling the need in the art for methods and reagents which employ fragments of differentially expressed genes of known, unknown (or previously unrecognized) function or consequence to provide diagnostic and therapeutic methods and reagents for diagnosis
15 and treatment of disease or infection, which conditions are characterized by such genes and gene products. It should be appreciated that it is the expression differences that are diagnostic of the altered state (e.g., predisease, disease, pathogenic, progression or infectious). Such genes associated with the altered state are likely to be the targets of drug discovery, whether the genes are the cause or the effect of the
20 condition, identification of such genes provides insight into which gene expression needs to be re-altered in order to reestablished the healthy state.

Summary of the Invention

 In one aspect, the invention provides methods for identifying gene(s)
25 which are differentially expressed, for example, in a normal healthy organism and an organism having a disease. The method involves producing and comparing hybridization patterns formed between samples of expressed mRNA or cDNA polynucleotide sequences obtained from either analogous cells, tissues or organs of a healthy organism and a diseased organism and a defined set of
30 oligonucleotide/polynucleotide/polynucleotide sequence probes from either an healthy organism or a diseased organism immobilized on a support. Those defined oligonucleotide/polynucleotide sequences are representative of the total expressed genetic component of the cells, tissues, organs or organism as defined the collection of partial cDNA sequences (ESTs). The differences between the hybridization
35 patterns permit identification of those particular EST or gene-specific oligonucleotide/polynucleotide sequences associated with differential expression, and the identification of the EST permits identification of the clone from which it was

derived and using ordinary skill further cloning and, if desired, sequencing of the full-length cDNA and genomic counterpart, i.e., gene, from which it was obtained.

5 In another aspect, the invention provides methods substantially similar to those described above, but which permit identification of those gene(s) of a pathogen which are expressed in any biological sample of an infected organism based on comparative hybridization of RNA/cDNA samples derived from a healthy versus infected organism, hybridized to an oligonucleotide/polynucleotide set representative of the gene coding complement of the pathogen of interest.

10 In another aspect, the invention provides methods substantially similar to those described above, but which permit identification of those ESTs-specific oligonucleotide/polynucleotide sequences of host gene(s) which represent genes being differentially expressed/ altered in expression by the disease state, or infection and are expressed in any biological sample of an infected organism based on comparative hybridization of RNA/cDNA samples derived from a healthy versus infected organism of interest.

15 In a further aspect, the methods described above and in detail below, also provide methods for diagnosis of diseases or infections characterized by differentially expressed genes, the expression of which has been altered as a result of infection by the pathogen or disease causing agent in question. All identified differences provide the basis for diagnostic testing be it the altered expression of endogenous genes or the patterned expression of the genes of the infecting organism. Such patterns of altered expression are defined by comparing RNA/cDNA from the two states hybridized against a panel of oligonucleotide/polynucleotides representing the expressed gene component of a cell, tissue, organ or organism as defined by its collection of ESTs.

20 Yet a further aspect of this invention provides a composition suitable for use in hybridization, which comprises a solid surface on which is immobilized at pre-defined regions thereon a plurality of defined oligonucleotide/polynucleotide sequences for hybridization, each sequence comprising a fragment of an EST isolated from a cDNA or DNA library prepared from at least one selected tissue or cell sample of a healthy (i.e., pre-disease state) animal, at least one analogous sample of an animal having a disease, at least one analogous sample of an animal infected with a pathogen or the pathogen itself, or any combination or multiple combinations thereof.

30 An additional aspect of the invention provides an isolated gene sequence which is differentially expressed in a normal healthy animal and an animal having a disease, and is identified by the methods above. Similarly, an isolated pathogen gene sequence which is expressed in tissue or cell samples of an infected animal can be identified by the methods above.

Yet another aspect of the invention is that it provides not only a means for a static diagnostic but also provides a means for a carrying out the procedure over time to measure disease progression as well as monitoring the efficacy of disease treatment regimes including an toxicological effects thereof.

5 Another aspect of the invention is an isolated protein produced by expression of the gene sequences identified above. Such proteins are useful in therapeutic compositions or diagnostic compositions, or as targets for drug development.

10 Other aspects and advantages of the present invention are described further in the following detailed description of the preferred embodiments thereof.

Detailed Description of the Invention

15 The present invention meets the unfulfilled needs in the art by providing methods for the identification and use of gene fragments and genes, even those of unknown full length sequence and unknown function, which are differentially expressed in a healthy animal and in an animal having a specific disease or infection by use of ESTs derived from DNA libraries of healthy and/or diseased/infected animals. Employing the methods of this invention permits the resulting identification and isolation of such genes by using their corresponding ESTs
20 and thereby also permits the production of protein products encoded by such genes. The genes themselves and/or protein products, if desired, may be employed in the diagnosis or therapy of the disease or infection with which the genes are associated and in the development of new drugs therefor.

25 It has been appreciated that one or more differentially identified EST or gene-specific oligonucleotide/polynucleotides define a pattern of differentially expressed genes diagnostic of a predisease, disease or infective state. A knowledge of the specific biological function of the EST is not required only that the ESTs identifies a gene or genes whose altered expression is associated reproducibly with the predisease, disease or infectious state. The differences permit the identification of
30 gene products altered in their expression by the disease and represent those products most likely to be targets of therapeutic intervention. Similarly, the product may be of the infecting organism itself and also be an effective target of intervention.

1. Definitions.

35 Several words and phrases used throughout this specification are defined as follows:

As used herein, the term "gene" refers to the genomic nucleotide sequence from which a cDNA sequence is derived, which cDNA produces an EST, as

described below. The term gene classically refers to the genomic sequence, which, upon processing, can produce different cDNAs, e.g., by splicing events. However, for ease of reading, any full-length counterpart cDNA sequence which gives rise to an EST will also be referred to by shorthand herein as a 'gene'.

5 The term "organism" includes without limitation, microbes, plants and animals.

 The term "animal" is used in its broadest sense to include all members of the animal kingdom, including humans. It should be understood, however, that according to this invention the same species of animal which provides the biological
10 sample also is the source of the defined immobilized oligonucleotide/polynucleotides as defined below.

 The term "pathogen" is defined herein as any molecule or organism which is capable of infecting an animal or plant and replicating its nucleic acid sequences in the cells or tissues of that animal or plant. Such a pathogen is generally
15 associated with a disease condition in the infected animal or plant. Such pathogens may include viruses, which replicate intra- or extra-cellularly, or other organisms, such as bacteria, fungi or parasites, which generally infect tissues or the blood. Certain pathogens or microorganisms are known to exist in sequential and distinguishable stages of development, e.g., latent stages, infective stages, and stages
20 which cause symptomatic diseases. In these different stages, the pathogens are anticipated to express differentially certain genes and/or turn on or off host cell gene expression.

 As used herein, the term "disease" or "disease state" refers to any condition which deviates from a normal or standardized healthy state in an organism
25 of the same species in terms of differential expression of the organism's genes. In other words, a disease state can be any illness or disorder be it of genetic or environmental origin, for example, an inherited disorder such as certain breast cancers, or a disorder which is characterized by expression of gene(s) normally in an inactive, 'turned off' state in a healthy animal, or a disorder which is characterized by
30 under-expression or no expression of gene(s) which is normally activated or 'turned on' in a normal healthy animal. Such differential expression of genes may also be detected in a condition caused by infection, inflammation, or allergy, a condition caused by development or aging of the animal, a condition caused by administration of a drug or exposure of the animal to another agent, e.g., nutrition, which affects
35 gene expression. Essentially, the methods described herein can be adapted to detect differential gene expression resulting from any cause, by manipulation of the defined oligonucleotide/polynucleotides and the samples tested as described below. The

concept of disease or disease state also includes its temporal aspects in terms of progression and treatment.

The phrase "differentially expressed" refers to those situations in which a gene transcript is found in differing numbers of copies, or in activated vs inactivated states, in different cell types or tissue types of an organism, having a selected disease as contrasted to the levels of the gene transcript found in the same cells or tissues of a healthy organism. Genes may be differentially expressed in differing states of activation in microorganisms or pathogens in different stages of development. For example, multiple copies of gene transcripts may be found in an organism having a selected disease, while only one, or significantly fewer copies, of the same gene transcript are found in a healthy organism, or vice-versa.

As used herein, the term "solid support" refers to any known substrate which is useful for the immobilization of large numbers of oligonucleotide/polynucleotide sequences by any available method to enable detectable hybridization of the immobilized oligonucleotide/polynucleotide sequences with other polynucleotide sequences in a sample. Among a number of available solid supports, one desirable example is the supports described in International Patent Application No. WO91/07087, published May 30, 1991. Also useful are supports such as but not limited to nitrocellulose, myelin, glass, silica and Pall Biodyne C[®]. It is also anticipated that improvements yet to be made to conventional solid supports may also be employed in this invention.

The term "surface" means any generally two-dimensional structure on a solid support to which the desired oligonucleotide/polynucleotide sequence is attached or immobilized. A surface may have steps, ridges, kinks, terraces and the like.

As used herein, the term "predefined region" refers to a localized area on a surface of a solid support on which is immobilized one or multiple copies of a particular oligonucleotide/polynucleotide sequence and which enables the identification of the oligonucleotide/polynucleotide at the position, if hybridization of that oligonucleotide/polynucleotide to a sample polynucleotide occurs.

By "immobilized" refers to the attachment of the oligonucleotide/polynucleotide to the solid support. Means of immobilization are known and conventional to those of skill in the art, and may depend on the type of support being used.

By "EST" or "Expressed Sequence Tag" is meant a partial DNA or cDNA sequence of about 150 to 500, more preferably about 300, sequential nucleotides of a longer sequence obtained from a genomic or cDNA library prepared from a selected cell, cell type, tissue or tissue type, organ or organism which longer

sequence corresponds to an mRNA of a gene found in that library. An EST is generally DNA. One or more libraries made from a single tissue type typically provide at least about 3000 different (i.e., unique) ESTs and potentially the full complement of all possible ESTs representing all cDNAs e.g., 50,000-100,000 in an animal such as a human. Further background and information on the construction of ESTs is described in M. D. Adams et al, Science, 252:1651-1656 (1991); and International Application Number PCT/US92/05222 (January 7, 1993).

As used herein, the term "defined oligonucleotide/polynucleotide sequence" refers to a known nucleotide sequence fragment of a selected EST or gene. This term is used interchangeably with the term "fragments of EST". These sequential sequences are generally comprised of between about 15 to about 45 nucleotides and more preferably between about 20 to about 25 nucleotides in length. Thus any single EST of 300 nucleotides in length may provide about 280 different defined oligonucleotide/polynucleotide sequences of 20 nucleotides in length (e.g., 20-mers). The lengths of the defined oligonucleotide/polynucleotides may be readily increased or decreased as desired or needed, depending on the limitations of the solid support on which they may be immobilized or the requirements of the hybridization conditions to be employed. The length is generally guided by the principle that it should be of sufficient length to insure that it is one average only represented once in the population to be examined. Generally, these defined oligonucleotide/polynucleotides are RNA or DNA and are preferably derived from the anti-sense strand of the EST sequence or from a corresponding mRNA sequence to enable their hybridization with samples of RNA or DNA. Modified nucleotides may be incorporated to increase stability and hybridization properties.

By the term "plurality of defined oligonucleotide/polynucleotide sequences" is meant the following. A surface of a solid support may immobilize a large number of "defined oligonucleotide/polynucleotides". For example, depending upon the nature of the surface, it can immobilize from about 300 to upwards of 60,000 defined 20-mer oligonucleotide/polynucleotides. It is anticipated that future improvements to solid surfaces will permit considerably larger such pluralities to be immobilized on a single surface. A "plurality" of sequences refers to the use on any one solid support of multiple different defined oligonucleotide/polynucleotides from a single EST from a selected library, as well as multiple different defined oligonucleotide/polynucleotides from different ESTs from the same library or many libraries from the same or different tissues, and may also include multiple identical copies of defined oligonucleotide/polynucleotides. Ultimately a plurality has at least one oligonucleotide/polynucleotide per expressed gene in the entire organism. For example, from a library producing about 5,000-10,000 ESTs, a single support can

include at least about 1-20 defined oligonucleotide/polynucleotides representing every EST in that library. The composition of defined oligonucleotide/polynucleotides which make up a surface according to this invention may be selected or designed as desired.

5 The term "sample" is employed in the description of this invention in several important ways. As used herein, the term "sample" encompasses any cell or tissue from an organism. Any desired cell or tissue type in any desired state may be selected to form a sample. For example, the sample cell desired may be a human T cell; the desired cell type for use in this invention may be a quiescent T cell or an
10 activated T cell.

 By the phrase "analogous sample" or "analogous cell or tissue" is meant that according to this invention when the ESTs which provide the defined oligonucleotide/polynucleotides are produced from a cDNA library prepared from a single tissue or cell type source sample, e.g., liver tissue of a human, then the samples
15 used to hybridize to those immobilized defined oligonucleotide/polynucleotides are preferably provided by the same type of sample from either a healthy or diseased animal, i.e., liver tissue of a healthy human and liver tissue of a diseased or infected human or from a human suspected of having that disease or infection. Alternatively, if the surface contains defined oligonucleotide/polynucleotides from multiple cells or
20 tissues, then the "samples" which are hybridized thereto can be but are not limited to samples obtained from analogous multiple tissues or cells.

 By the term "detectably hybridizing" means that the sample from the healthy organism or diseased or infected organism is contacted with the defined oligonucleotide/polynucleotides on the surface for sufficient time to permit the
25 formation of patterns of hybridization on the surfaces caused by hybridization between certain polynucleotide sequences in the samples with the certain immobilized defined oligonucleotide/polynucleotides. These patterns are made detectable by the use of available conventional techniques, such as fluorescent labelling of the samples. Preferably hybridization takes place under stringent conditions, e.g., revealing
30 homologies of about 95%. However, if desired, other less stringent conditions may be selected. Techniques and conditions for hybridization at selected stringencies are well known in the art [see, e.g., Sambrook et al, Molecular Cloning. A Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY (1989)].

35 II. Compositions of The Invention

 The present invention is based upon the use of ESTs from any desired cell or tissue in known technologies for oligonucleotide/polynucleotide hybridization.

A. ESTs

An EST, as defined above, is for an animal, a sequence from a cDNA clone that corresponds to an mRNA. The EST sequences useful in the present invention are isolated preferably from cDNA libraries using a rapid screening and sequencing technique. Custom made cDNA libraries are made using known techniques. See, generally, Sambrook et al, cited above. Briefly, mRNA from a selected cell or tissue is reverse transcribed into complementary DNA (cDNA) using the reverse transcriptase enzyme and made double-stranded using RNase H coupled with DNA polymerase or reverse transcriptase. Restriction enzyme sites are added to the cDNA and it is cloned into a vector. The result is a cDNA library. Alternatively, commercially available cDNA libraries may be used. Libraries of cDNA can also be generated from recombinant expression of genomic DNA using known techniques, including polymerase chain reaction-derived techniques.

ESTs (which can range from about 150 to about 500 nucleotides in length, preferably about 300 nucleotides) can be obtained through sequence analysis from either end of the cDNA insert. Desirably, the DNA libraries used to obtain ESTs use directional cloning methods so that either the 5' end of the cDNA (likely to contain coding sequence) or the 3' end (likely to be a non-coding sequence) can be selectively obtained.

In general, the method for obtaining ESTs comprises applying conventional automated DNA sequencing technology to screen clones, advantageously randomly selected clones, from a cDNA library. The cDNA libraries from the desired tissue can be preprocessed, or edited, by conventional techniques to reduce repeated sequencing of high and intermediate abundance clones and to maximize the chances of finding rare messages from specific cell populations. Preferably, preprocessing includes the use of defined composition prescreening probes, e.g., cDNA corresponding to mitochondria, abundant sequences, ribosomes, actins, myelin basic polypeptides, or any other known high abundance peptide. These prescreening probes used for preprocessing are generally derived from known ESTs. Other useful preprocessing techniques include subtraction hybridization, which preferentially reduces the population of highly represented sequences in the library [e.g., see Fargnoli et al, Anal. Biochem., 187:364 (1990)] and normalization, which results in all sequences being represented in approximately equal proportions in the library [Patanjali et al, Proc. Natl. Acad. Sci. USA, 88:1943 (1991)]. Additional prescreening/differential screening approaches are known to those skilled in the art.

ESTs can then be generated from partial DNA sequencing of the selected clones. The ESTs useful in the present invention are preferably generated using low redundancy of sequencing, typically a single sequencing reaction. While

single sequencing reactions may have an accuracy as low as 90%, this nevertheless provides sufficient fidelity for identification of the sequence and design of PCR primers.

If desired, the location of an EST in a full length cDNA is determined by analyzing the EST for the presence of coding sequence. A conventional computer program is used to predict the extent and orientation of the coding region of a sequence (using all six reading frames). Based on this information, it is possible to infer the presence of start or stop codons within a sequence and whether the sequence is completely coding or completely non-coding or a combination of the two. If start or stop codons are present, then the EST can cover both part of the 5'-untranslated or 3'-untranslated part of the mRNA (respectively) as well as part of the coding sequence. If no coding sequence is present, it is likely that the EST is derived from the 3' untranslated sequence due to its longer length and the fact that most cDNA library construction methods are biased toward the 3' end of the mRNA. It should be understood that both coding and non-coding regions may provide ESTs equally useful in the described invention.

A number of specific ESTs suitable for use in the present invention are described above Adams et al (*supra*), which may be incorporated by reference herein, to describe non-essential examples of desirable ESTs. Other ESTs exist in the art which may also be useful in this invention, as will ESTs yet to be developed by these known techniques.

B. Preparing the Solid Support of the Invention

Oligonucleotide sequences which are fragments of defined sequence are derived from each EST by conventional means, e.g., conventional chemical synthesis or recombinant techniques. Each defined oligonucleotide/polynucleotide sequence as described above is a fragment, can be, but is not necessarily an anti-sense fragment, of an EST isolated from a DNA library prepared from a selected cell or tissue type from a selected animal. For use in the present invention, it is presently preferred that the defined oligonucleotide/polynucleotide sequences are 20-25mers. As described above, for each EST a number of such 20-25mers may be generated. The lengths may vary as described above as well as the composition. For example oligonucleotide/polynucleotides can be modified based on the Oligo 4.0 or similar programs to predict hybridization potential or to include modified nucleotides for the reasons given above. It is also appreciated that large DNA segments may be employed including entire ESTs or even full length genes particular when inserted into cloning vectors.

A plurality of these defined oligonucleotide/polynucleotide sequences are then attached to a selected solid support conventionally used for the attachment of nucleotide sequences again by known means. In contrast to other technologies available in the art, this support is designed to contain defined, not random, oligonucleotide/polynucleotide sequences. The EST fragments, or defined oligonucleotide/polynucleotide sequences, immobilized on the solid support can include fragments of one or more ESTs from a library of at least one selected tissue or cell sample of a healthy animal, at least one analogous sample of the animal having a disease, at least one analogous sample of the animal infected with a pathogen, and any combination thereof.

Numerous conventional methods are employed for attaching biological molecules such as oligonucleotide/polynucleotide sequences to surfaces of a variety of solid supports. See, e.g., Affinity Techniques, Enzyme Purification: Part B, Methods in Enzymology, Vol. 34, ed. W.B. Jakoby, M. Wilcheck, Acad. Press, NY (1974); Immobilized Biochemicals and Affinity Chromatography, Advances in Experimental Medicine and Biology, vol. 42, ed. R. Dunlap, Plenum Press, NY (1974); U. S. Patent No. 4,762,881; U. S. Patent No. 4,542,102; European Patent Publication No. 391,608 (October 10, 1990); U. S. Patent No. 4,992,127 (Nov. 21, 1989).

One desirable method for attaching oligonucleotide/polynucleotide sequences derived from ESTs to a solid support is described in International Application No. PCT/US90/06607 (published May 30, 1991). Briefly, this method involves forming predefined regions on a surface of a solid support, where the predefined regions are capable of immobilizing ESTs. The methods make use of binding substances attached to the surface which enable selective activation of the predefined regions. Upon activation, these binding substances become capable of binding and immobilizing oligonucleotide/polynucleotides based on EST or longer gene sequences.

Any of the known solid substrates suitable for binding oligonucleotide/polynucleotides at pre-defined regions on the surface thereof for hybridization and methods for attaching the oligonucleotide/polynucleotides thereto may be employed by one of skill in the art according to this invention. Similarly, known conventional methods for making hybridization of the immobilized oligonucleotide/polynucleotides detectable, e.g., fluorescence, radioactivity, photoactivation, biotinylation, solid state circuitry, and the like may be used in this invention.

Thus, by resorting to known techniques, the invention provides a composition suitable for use in hybridization which consists of a surface of a solid

support on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences for hybridization. For example, one composition of this invention is a solid support on which are immobilized oligos of EST fragments from a library constructed from a single cell type, e.g., a human stem cell, or a single tissue, e.g., human liver, from a healthy human. Still another composition of this invention is another solid support on which are immobilized oligos of EST fragments from a library constructed from a single cell type or a tissue from a human having a selected disease or predisposition to a selected disease, e.g., liver cancer.

Another embodiment of the compositions of this invention include a single solid support having oligonucleotides of ESTs from both single cell or single tissue libraries from both a healthy and diseased human. Still other embodiments include a single support on which are immobilized oligos of EST fragments from more than one tissue or cell library from a healthy human or a single support on which are immobilized more than one tissue or cell library from both healthy and diseased animals or humans. A preferred composition of this invention is anticipated to be a single support containing oligos of ESTs for all known cells and tissues from a selected organism.

III. The Methods of the Invention

A. Identification of Genes

The present invention employs the compositions described above in methods for identifying genes which are differentially expressed in a normal healthy organism and an organism having a disease or infection. These methods may be employed to detect such genes, regardless of the state of knowledge about the function of the gene. The method of this invention by use of the compositions containing multiple defined EST fragments from a single gene as described above is able to detect levels of expression of genes or in other cases simply the expression or lack thereof, which differ between normal, healthy organisms and organisms having a selected disease, disorder or infection.

One such method employs a first surface of a solid support on which is immobilized at pre-defined regions thereon a plurality of defined oligonucleotide/polynucleotide sequences, described above, of EST or longer gene fragment isolated from a cDNA library prepared from at least one selected tissue or cell sample of a healthy animal (the "healthy test surface") and a second such surface on which is immobilized at pre-defined regions a plurality of defined oligonucleotide/polynucleotide sequences of EST or longer gene fragment isolated from at least one analogous tissue of an animal having a selected disease (the "disease

test surface"). These test surfaces may be standardized for the selected animal or selected cell or tissue sample from that animal (i.e., they are prescreened for polymorphisms in the species population).

Polynucleotide sequences are then isolated from mRNA and/or
5 cDNA from a biological sample from a known healthy animal ("healthy control") and a second sample is similarly prepared from a sample from a known diseased animal ("disease sample"). These two samples are desirably selected from the cell or tissue analogous to that which provided the immobilized oligonucleotide/polynucleotides.

According to the method the healthy control sample is
10 contacted with one set of the healthy test surface and the disease test surface described above for a time sufficient to permit detectable hybridization to occur between the sample and the immobilized defined oligonucleotide/polynucleotides on each surface. The results of this hybridization are a first hybridization pattern formed between the nucleotides of healthy control and the healthy test surface and a second
15 hybridization pattern formed between the nucleotides of healthy control sample and the disease test surface.

In a similar manner, the disease sample is detectably hybridized to another set of healthy test and disease test surfaces, forming a third hybridization pattern between the disease sample and healthy test surface and a fourth hybridization
20 pattern between the disease sample and the disease test surface.

Comparing the four hybridization patterns permits detection of those defined oligonucleotide/polynucleotides which are differentially expressed between the healthy control and the disease sample by the presence of differences in the hybridization patterns at pre-defined regions. The
25 oligonucleotide/polynucleotides on each surface which correspond to the pattern differences may be readily identified with the corresponding EST or longer gene fragment from which the oligonucleotide/polynucleotides are obtained.

In another embodiment of the method of this invention, the same process is employed, with the exception that plurality of defined
30 oligonucleotide/polynucleotide sequences forming the healthy test sample and the disease test sample surfaces are immobilized on a single solid support. For example, each fragment of an EST or longer gene fragment on the surface is isolated from at least two cDNA libraries prepared from a selected cell or tissue sample of a healthy animal and an analogous selected cell or tissue sample of an animal having a disease.

35 According to this embodiment, the healthy control sample is detectably hybridized to a copy of this single solid surface, forming one hybridization pattern with oligonucleotide/polynucleotides associated with both the healthy and diseased animal. Similarly, the disease sample is detectably hybridized to a second

copy of this single solid surface, forming one hybridization pattern with oligonucleotide/polynucleotides associated with both the healthy and diseased animal.

Comparing the two hybridization patterns permits detection of those defined oligonucleotide/polynucleotides which are differentially expressed between the healthy control and the disease sample by the presence of differences in the hybridization patterns at pre-defined regions. The oligonucleotide/polynucleotides on each surface which correspond to the pattern differences may be readily identified with the corresponding EST or longer gene fragment from which the oligonucleotide/polynucleotides are obtained.

The identification of one or more ESTs as the source of the defined oligonucleotide/polynucleotide which produced a "difference" in hybridization patterns according to these methods permits ready identification of the gene from which those ESTs were derived. Because oligonucleotides are of sufficient length that they will hybridize under stringent conditions only with a RNA/cDNA for that gene to which they correspond, the oligo can be used to identify the EST and in turn the clone from which it was derived and by subsequent cloning, obtain the sequence of the full-length cDNA and its genomic counterparts, i.e., the gene, from which it was obtained.

In other words, the ESTs identified by the method of this invention can be employed to determine the complete sequence of the mRNA, in the form of transcribed cDNA, by using the EST as a probe to identify a cDNA clone corresponding to a full-length transcript, followed by sequencing of that clone. The EST or the full length cDNA clone can also be used as a probe to identify a genomic clone or clones that contain the complete gene including regulatory and promoter regions, exons, and introns.

It should be appreciated that one does not have to be restricted in using ESTs from a particular tissue from which probe RNA or cDNA is obtained, rather any or all ESTs (known or unknown) may be placed on the support. Hybridization will be used to form diagnostic patterns or to identify which particular EST is detected. For example, all known ESTs from an organism are used to produce a "master" solid support to which control sample and disease samples are alternately hybridized. One then detects a pattern of hybridization associated with the particular disease state which then forms the basis of a diagnostic test or the isolation of disease specific ESTs from which the intact gene may be cloned and sequenced leading ultimately to a defined therapeutic target.

Methods for obtaining complete gene sequences from ESTs are well-known to those of skill in the art. See, generally, Sambrook et al, cited above. Briefly, one suitable method involves purifying the DNA from the clone that was

sequenced to give the EST and labeling the isolated insert DNA. Suitable labeling systems are well known to those of skill in the art [see, eg. Basic Methods in Molecular Biology, L. G. Davis et al, ed., Elsevier Press, NY (1986)]. The labeled EST insert is then used as a probe to screen a lambda phage cDNA library or a plasmid cDNA library, identifying colonies containing clones related to the probe cDNA which can be purified by known methods. The ends of the newly purified clones are then sequenced to identify full length sequences and complete sequencing of full length clones is performed by enzymatic digestion or primer walking. A similar screening and clone selection approach can be applied to clones from a genomic DNA library.

Additionally, an EST or gene identified by this method as associated with inherited disorders can be used to determine at what stage during embryonic development the selected gene from which it is derived is developed by screening embryonic DNA libraries from various stages of development, e.g. 2-cell, 8-cell, etc., for the selected gene. As has been mentioned above, the invention may be applied in additional temporal modes for monitoring the progression of a disease state, the efficacy of a particular treatment modality or the aging process of an individual.

Thus, the methods of this invention permit the identification, isolation and sequencing of a gene which is differentially expressed in a selected disease/infection. As described in more detail below, the identified gene may then be employed to obtain any protein encoded thereby, or may be employed as a target for diagnostic methods or therapeutic approaches to the treatment of the disease, including, e.g., drug development.

The same methods as described above for the identification of genes, including genes of unknown function, which are differentially expressed in a disease state, may also be employed to identify other genes of interest. For example, another embodiment of this invention includes a method for identifying a gene of a pathogen which is expressed in a biological sample of an animal infected with that pathogen or the gene of the host which is altered in its expression as a result of the infection.

One such method employs a healthy test surface as described above, employing defined oligonucleotide/polynucleotides from a sample of a healthy, uninfected animal. The second such surface has immobilized at pre-defined regions thereon a plurality of defined oligonucleotide/polynucleotide sequences of ESTs isolated from at least one analogous tissue or cell sample of an infected animal (the "infection test surface"). Polynucleotide sequences are isolated from a biological sample from a healthy animal ("healthy control") and a second sample is similarly

prepared from an animal infected with the selected pathogen ("infection sample"). These two samples are desirably selected from the cell or tissue analogous to that which provided the immobilized oligonucleotide/polynucleotides. It would also be possible to provide samples from the nucleic acid of the pathogen itself.

5 According to the method the healthy control sample is contacted with one set of the healthy test surface and the infection test surface described above for a time sufficient to permit detectable hybridization to occur between the sample and the immobilized defined oligonucleotide/polynucleotides on each surface. The results of this hybridization are a first hybridization pattern formed
10 between the nucleotides of healthy control and the healthy test surface and a second hybridization pattern formed between the nucleotides of healthy control sample and the infection test surface.

 In a similar manner, the infection sample is detectably hybridized to another set of healthy test and infection test surfaces, forming a third
15 hybridization pattern between the infection sample and healthy test surface and a fourth hybridization pattern between the infection sample and the infection test surface.

 Comparing the four hybridization patterns permits detection of those defined oligonucleotide/polynucleotides which are differentially expressed
20 between the healthy animal and the animal infected with the pathogen by the presence of differences in the hybridization patterns at pre-defined regions. As mentioned differential expression is not required and simple qualitative analysis is possible by reference to gene expression which is simply present or absent.

 A second embodiment of this method parallels the second
25 embodiment of the method as applied to disease above, i.e., the same process is employed, with the exception that plurality of defined oligonucleotide/polynucleotide sequences forming the healthy test sample surface and the infection test sample surface are immobilized on a single solid support. The resulting first hybridization pattern (healthy control sample with healthy/infection test sample) and second
30 hybridization pattern (infection sample with healthy/infection test sample) permits detection of those defined oligonucleotide/polynucleotides which are differentially expressed between the healthy control and the infection sample by the presence of differences in the hybridization patterns at pre-defined regions. The oligonucleotide/polynucleotides on each surface which correspond to the pattern
35 differences may be readily identified with the corresponding ESTs from which the oligonucleotide/polynucleotides are obtained.

 As described above for the methods for identifying differential gene expression between diseased and healthy animals, the

oligonucleotide/polynucleotides on each surface which correspond to the pattern differences may be readily identified with the corresponding ESTs from which the oligonucleotide/polynucleotide sequences are obtained and the genes expressed by the pathogen identified for similar purposes. Other embodiments of these methods may be developed with resort to the teaching herein, by altering the samples which provide the defined oligonucleotide/polynucleotides. For example, an EST, identified with a differentially expressed gene by the method of this invention is also useful in detecting genes expressed in the various stages of an pathogen's development, particularly the infective stage and following the cours of drug treatment and emergence of resistant variants. For example, employing the techniques described above, the EST can be used for detecting a gene in various stages of the parasitic *Plasmodium* species life cycle, which include blood stages, liver stages, and gametocyte stages.

B. Diagnostic Methods

In addition to use of the methods and compositions of this invention for identifying differentially expressed genes, another embodiment of this invention provides diagnostic methods for diagnosing a selected disease state, or a selected state resulting from aging, exposure to drugs or infection in an animal. According to this aspect of the invention, a first surface, described as the healthy test surface above, and a second surface, described as the disease test surface or infection test surface, are prepared depending on the disease or infection to be diagnosed. The same processes of detectable hybridization to a first and second set of these surfaces with the healthy control sample and disease/infection sample are followed to provide the four above-described hybridization patterns, i.e., healthy control sample with healthy test surface; healthy control sample with disease/infection test surface; disease/infection sample with healthy test surface; and disease/infection sample with disease/infection test surface.

The diagnosis of disease or infection is provided by comparing the four hybridization patterns. Substantial differences between the first and third hybridization patterns, respectively, and the second and fourth hybridization patterns, respectively, indicate the presence of the selected disease or infection in said animal. Substantial similarities in the first and third hybridization patterns and second and fourth hybridization patterns indicates the absence of disease or infection.

A similar embodiment utilizes the single surface bearing both the healthy test surface defined oligonucleotide/polynucleotides and the disease/infection test surface defined oligonucleotide/polynucleotides as described above. Parallel process steps as described above for detection of genes differentially expressed in disease and infected states are followed, resulting in a first hybridization

pattern (healthy control sample with single healthy and disease/infection test sample) and a second hybridization pattern (disease/infection sample with another copy of the single healthy and disease/infection test sample).

5 Diagnosis is accomplished by comparing the two hybridization patterns, wherein substantial differences between the first and second hybridization patterns indicate the presence of the selected disease or infection in the animal being tested. Substantially similar first and second hybridization patterns indicate the absence of disease or infection. This like many of the foregoing embodiments may use known or unknown ESTs derived from many libraries.

10 C. *Other Methods of the Invention*

 As is obvious to one of skill in the art upon reading this disclosure, the compositions and methods of this invention may also be used for other similar purposes. For example, the general methods and compositions may be adapted easily by manipulation of the samples selected to provide the standardized
15 defined oligonucleotide/polynucleotides, and selection of the samples selected for hybridization thereto. One such modification is the use of this invention to identify cell markers of any type, e.g., markers of cancer cells, stem cell markers, and the like. Another modification involves the use of the method and compositions to generate hybridization patterns useful for forensic identification or an 'expression fingerprint'
20 of genes for identification of one member of a species from another. Similarly, the methods of this invention may be adapted for use in tissue matching for transplantation purposes as well as for molecular histology, i.e., to enable diagnosis of disease or disorders in pathology tissue samples such as biopsies. Still another use of this method is in monitoring the effects of development and aging upon the gene
25 expression in a selected animal, by preparing surfaces bearing oligonucleotide/polynucleotides prepared from samples of standardized younger members of the species being tested. Additionally the patient can serve as an internal control by virtue of having the method applied to blood samples every 5-10 years during his lifetime.

30 Still another intriguing use of this method is in the area of monitoring the effects of drugs on gene expression, both in laboratories and during clinical trials with animal, especially humans. Because the method can be readily adapted by altering the above parameters, it can essentially be employed to identify differentially expressed genes of any organism, at any stage of development, and
35 under the influence of any factor which can affect gene expression.

IV. *The Genes and Proteins Identified*

Application of the compositions and methods of this invention as above described also provide other compositions, such as any isolated gene sequence which is differentially expressed between a normal healthy animal and an animal having a disease or infection. Another embodiment of this invention is any isolated pathogen gene sequence which is expressed in tissue or cell samples of an infected animal. Similarly an embodiment of this invention is any gene sequence identified by the methods described herein.

These gene sequences may be employed in conventional methods to produce isolated proteins encoded thereby. To produce a protein of this invention, the DNA sequences of a desired gene identified by the use of the methods of this invention or portions thereof are inserted into a suitable expression system. Desirably, a recombinant molecule or vector is constructed in which the polynucleotide sequence encoding the protein is operably linked to a heterologous expression control sequence permitting expression of the human protein. Numerous types of appropriate expression vectors and host cell systems are known in the art for mammalian (including human) expression, insect, e.g., baculovirus expression, yeast, fungal, and bacterial expression, by standard molecular biology techniques.

The transfection of these vectors into appropriate host cells, whether mammalian, bacterial, fungal, or insect, or into appropriate viruses, can result in expression of the selected proteins. Suitable host cells or cell lines for transfection, and viruses, as well as methods for the construction and transfection of such host cells and viruses are well-known. Suitable methods for transfection, culture, amplification, screening, and product production and purification are also known in the art.

The genes and proteins identified by this invention can be employed, if desired in diagnostic compositions useful for the diagnosis of a disease or infection using conventional diagnostic assays. For example, a diagnostic reagent can be developed which detectably targets a gene sequence or protein of this invention in a biological sample of an animal. Such a reagent may be a complementary nucleotide sequence, an antibody (monoclonal, recombinant or polyclonal), or a chemically derived agonist or antagonist. Alternatively, the proteins and polynucleotide sequences of this invention, fragments of same, or complementary sequences thereto, may themselves be useful as diagnostic reagents for diagnosing disease states with which the ESTs of the invention are associated. These reagents may optionally be labelled using diagnostic labels, such as radioactive labels, colorimetric enzyme label systems and the like conventionally used in diagnostic or therapeutic methods, e.g., Northern and Western blotting, antigen-antibody binding and the like. The selection of the appropriate assay format and label system is within the skill of the art and may

readily be chosen without requiring additional explanation by resort to the wealth of art in the diagnostic area.

Additionally, genes and proteins identified according to this invention may be used therapeutically. For example, the EST-containing gene sequences may be useful in gene therapy, to provide a gene sequence which in a disease is not properly or sufficiently expressed. In such a method, a selected gene sequence of this invention is introduced into a suitable vector or other delivery system for delivery to a cell containing a defect in the selected gene. Suitable delivery systems are well known to those of skill in the art and enable the desired EST or gene to be incorporated into the target cell and to be translated by the cell. The EST or gene sequence may be introduced to mutate the existing gene by recombination or provide an active copy thereof in addition to the inactive gene to replace its function.

Alternatively, a protein encoded by an EST or gene of the invention may be useful as a therapeutic reagent for delivery of a biologically active protein, particularly when the disease state is associated with a deficiency of this protein. Such a protein may be incorporated into an appropriate therapeutic formulation, alone or in combination with other active ingredients. Methods of formulating such therapeutic compositions, as well as suitable pharmaceutical carriers, and the like, are well known to those of skill in the art. Still an additional method of delivering the missing protein encoded by an EST, or the gene from which a selected EST was derived, involves expressing it directly *in vivo*. Systems for such *in vivo* expression are well known in the art.

Yet another use of the ESTs, genes identified according to the methods of this invention, or the proteins encoded thereby is a target for the screening and development of natural or synthetic chemical compounds which have utility as therapeutic drugs for the treatment of disease states associated with the identified genes and ESTs derived therefrom. As one example, a compound capable of binding to such a protein encoded by such a gene and either preventing or enhancing its biological activity may be a useful drug component for the treatment or prevention of such disease states.

Conventional assays and techniques may be used for the screening and development of such drugs. As one example, a method for identifying compounds which specifically bind to or inhibit or activate proteins encoded by these gene sequences can include simply the steps of contacting a selected protein or gene product, with a test compound to permit binding of the test compound to the protein; and determining the amount of test compound, if any, which is bound to the protein. Such a method may involve the incubation of the test compound and the protein immobilized on a solid support. Still other conventional methods of drug screening

can involve employing a suitable computer program to determine compounds having similar or complementary chemical structures to that of the gene product or portions thereof and screening those compounds either for competitive binding to the protein to detect enhanced or decreased activity in the presence of the selected compound.

5 Thus, through use of such methods, the present invention is anticipated to provide compounds capable of interacting with these genes, ESTs, or encoded proteins, or fragments thereof, and either enhancing or decreasing the biological activity, as desired. Such compounds are believed to be encompassed by this invention.

10 Numerous modifications and variations of the present invention are included in the above-identified specification and are expected to be obvious to one of skill in the art. Such modifications and alterations to the compositions and processes of the present invention are believed to be encompassed in the scope of the claims appended hereto.

15

WHAT IS CLAIMED IS:

1. A method for identifying genes which are differentially expressed in two different pre-determined states of an organism comprising:

- 5 a. providing a first surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene, isolated from a DNA library prepared from at least one selected cell, tissue, organ or organism sample in a first
10 state and present in excess relative to the polynucleotide to be hybridized;
- b. providing a second surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene, isolated from a DNA library
15 prepared from at least one selected cell, tissue, organ or organism sample in a second state and present in excess relative to the polynucleotide to be hybridized;
- c. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from a said organism in said first state, said sample selected from sources analogous to the sources of step (a), said
20 hybridization sufficient to form a first and second hybridization pattern on each said first and second surface,
- d. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from said organism in said second state, said sample selected from sources analogous to the sources of step (c), said
25 hybridization sufficient to form a third and fourth hybridization pattern on each said first and second surface,
- e. comparing at least two of the four hybridization patterns, wherein genes differentially expressed in said first and second states are identified by the presence of differences in the hybridization patterns at pre-defined regions;
- 30 f. identifying the oligonucleotide/polynucleotides on each surface which correspond to said pattern differences and the corresponding ESTs or larger gene fragment from which the oligonucleotide/polynucleotides were obtained, whereby identification of the EST or larger gene fragment permits identification of the gene from which the ESTs or larger gene fragment were derived.

35

2. The method according to Claim 1 wherein said first and second states are respectively healthy and disease; pathogen uninfected and pathogen infected; a first progression state and a second progression of a disease or infection; a first treatment state and a second treatment state of a disease or infection; or a first developmental and a second developmental state.

3. The method according to Claim 1 wherein said organism is a plant or an animal.

4. The method according to Claim 3 wherein said animal is a human.

5. A method for identifying genes which are differentially expressed in a normal healthy animal and an animal having a disease comprising:

a. providing a first surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene, isolated from a DNA library prepared from at least one selected cell, tissue, organ or organism sample in a healthy animal and present in excess relative to the polynucleotide to be hybridized;

b. providing a second surface on which is immobilized at pre-defined regions of said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene, isolated from a DNA library prepared from at least one selected cell, tissue, organ or organism sample from an animal having said disease and present in excess relative to the polynucleotide to be hybridized;

c. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from a healthy animal, said sample selected from sources analogous to the sources of step (a), said hybridization sufficient to form a first and second hybridization pattern on each said first and second surface, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a first and second hybridization pattern on each said first and second surface;

d. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from an animal having said disease, said sample selected from a cell or tissue sample analogous to the sample of step (c), said hybridization sufficient to form a third and fourth hybridization pattern on each
5 said first and second surface,

e. comparing at least two of the four hybridization patterns, wherein genes differentially expressed in said first and second states are identified by the presence of differences in the hybridization patterns at pre-defined regions;

f. identifying the oligonucleotide/polynucleotides on each surface
10 which correspond to said pattern differences and the corresponding ESTs or larger gene fragment from which the oligonucleotide/polynucleotides were obtained, whereby identification of the EST or larger gene fragment permits identification of the gene from which the ESTs or larger gene fragment were derived.

15 6. A method for identifying genes which are differentially expressed in a normal healthy animal and an animal having a disease comprising:

a. providing a surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST,
20 an entire EST a fragment of a gene or an entire gene isolated from a DNA library prepared from the group selected from at least one selected cell, tissue, organ or organism sample in of a healthy animal and an analogous selected sample of an animal having said disease and both present in excess relative to the polynucleotide to be hybridized;

25 b. detectably hybridizing to a first copy of said surface polynucleotide sequences isolated from a healthy animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a first hybridization pattern on said surface;

c. detectably hybridizing to a second copy of said surface
30 polynucleotide sequences isolated from an animal having said disease, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a second hybridization pattern on said surface;

d. comparing the two hybridization patterns, wherein genes differentially expressed in a disease state are identified by the presence of differences
35 in the hybridization patterns at pre-defined regions;

e. identifying the oligonucleotide/polynucleotides on each surface which correspond to said pattern differences and the corresponding ESTs from which the oligonucleotide/polynucleotides are obtained, whereby identification of the EST permits identification of the gene from which the ESTs were derived.

5

7. A method for identifying a gene of a pathogen which is expressed in a biological sample of an animal infected with said pathogen comprising:

a. providing a first surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene isolated from a DNA library prepared from at least one selected cell, tissue, organ or organism sample of a healthy, uninfected animal and present in excess relative to the polynucleotide to be hybridized;

15

b. providing a second surface on which is immobilized at pre-defined regions of said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene isolated from at least one selected cell, tissue, organ or organism sample of an infected animal;

20

c. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from a healthy animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form first and second hybridization patterns on each said first and second surface,

25

d. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from an infected animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form third and fourth hybridization patterns on each said first and second surface,

30

e. comparing the four hybridization patterns, wherein genes of said pathogen which are expressed in an infected animal are identified by the presence of differences in the hybridization patterns at pre-defined regions;

f. identifying the oligonucleotide/polynucleotides on each surface which correspond to said pattern differences and the corresponding ESTs from which the oligonucleotide/polynucleotides are obtained, whereby identification of the EST permits identification of the gene from which the ESTs were derived.

35

8. A method for identifying a gene of a pathogen which is expressed in a biological sample of an animal infected with said pathogen comprising:

- a. providing a surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene isolated from a DNA library prepared from the group selected from at least one selected cell, tissue, organ or organism sample in of a healthy animal and an analogous selected sample of an animal having said disease and both present in excess relative to the polynucleotide to be hybridized
- b. detectably hybridizing to a first copy of said surface polynucleotide sequences isolated from a sample from a healthy animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a first hybridization pattern on said surface;
- c. detectably hybridizing to a second copy of said surface polynucleotide sequences isolated from a sample from an infected animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a second hybridization pattern on said surface;
- d. comparing the two hybridization patterns, wherein genes of said pathogen which are expressed in an infected animal are identified by the presence of differences in the hybridization patterns at pre-defined regions;
- e. identifying the oligonucleotide/polynucleotides on each surface which correspond to said pattern differences and the corresponding ESTs from which the oligonucleotide/polynucleotides are obtained, whereby identification of the EST permits identification of the gene from which the ESTs were derived.

9. A composition suitable for use in hybridization comprising a solid surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences for hybridization, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene isolated from a DNA library prepared from the group selected from at least one selected cell, tissue, organ or organism sample of a healthy animal, at least one analogous sample of said animal having a disease, at least one analogous sample of said animal infected with a microbial pathogen, and any combination thereof.

10. An isolated gene sequence which is differentially expressed in a normal healthy animal and an animal having a disease, identified by the method of claim 1.

5 11. An isolated pathogen gene sequence which is expressed in tissue or cell samples of an infected animal identified by the method of claim 7.

12. A diagnostic composition useful for the diagnosis of a disease comprising a reagent capable of detectably targeting a gene sequence of claim 10 in a
10 biological sample of an animal.

13. A diagnostic composition useful for the diagnosis of infection by a pathogen comprising a reagent capable of detectably targeting a gene sequence of claim 11 in a biological sample of an animal.

15 14. An isolated protein produced by expression of a gene sequence of claim 10.

15. An isolated pathogen protein produced by expression of a gene
20 sequence of claim 11.

16. A therapeutic composition comprising a protein or fragment thereof selected from the group consisting of a protein of claim 10 and a protein of claim 15.

25 17. A method for diagnosing a selected disease or infection in an animal comprising:

a. providing a first surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST,
30 an entire EST a fragment of a gene or an entire gene, isolated from a DNA library prepared from at least one selected cell, tissue, organ or organism sample of a healthy animal and present in excess relative to the polynucleotide to be hybridized;

b. providing a second surface on which is immobilized at pre-defined regions of said surface a plurality of defined oligonucleotide/polynucleotide
35 sequences, each sequence comprising a fragment of an EST isolated from at least one said tissue of an animal having said disease;

5 c. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a DNA library prepared from a sample from a healthy animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a first and second hybridization pattern on each said first and second surface;

10 d. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a DNA library prepared from a sample from an animal having said disease, said sample selected from a cell or tissue sample analogous to the sample of step (c), said hybridization sufficient to form a third and fourth hybridization pattern on each said first and second surface;

15 e. comparing the four hybridization patterns, wherein substantial differences between the first and third hybridization patterns and the second and fourth hybridization patterns indicates the presence of said selected disease or infection in said animal, and substantial similarities in said first and third hybridization patterns and second and fourth hybridization patterns indicates the absence of disease or infection.

18. A method for diagnosing a selected disease or infection in an animal comprising:

20 a. providing a surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence comprising a fragment of an EST isolated from a DNA library prepared from the group consisting of a selected cell or tissue sample of a healthy animal and an analogous selected cell or tissue sample of an animal having said disease;

b. detectably hybridizing to a first copy of said surface polynucleotide sequences isolated from a sample from a healthy animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a first hybridization pattern on said surface;

30 c. detectably hybridizing to a second copy of said surface polynucleotide sequences isolated from a DNA library prepared from a sample from an animal having said disease, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a second hybridization pattern on said surface;

35 d. comparing the two hybridization patterns, wherein substantial differences between the first and second hybridization patterns indicates the presence of said selected disease or infection in said animal, and substantial similarities in said first and second hybridization patterns indicates the absence of disease or infection.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US95/01863

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) : C12Q 1/68

US CL : 435/6

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS, CAS, BIOSIS

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	ANALYTICAL BIOCHEMISTRY, VOLUME 187, ISSUED 1990, FARGNOLI ET AL, "LOW-RATIO HYBRIDIZATION SUBTRACTION", PAGES 364-373, SEE ENTIRE DOCUMENT.	1-18
Y	PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES USA, VOLUME 88, ISSUED MARCH 1991, PATANJALI ET AL, "CONSTRUCTION OF A UNIFORM-ABUNDANCE (NORMALIZED) CDNA LIBRARY", PAGES 1943-1947, SEE ENTIRE DOCUMENT.	1-18
Y	SCIENCE, VOLUME 245, ISSUED 29 SEPTEMBER 1989, OLSON ET AL. "A COMMON LANGUAGE FOR PHYSICAL MAPPING OF THE HUMAN GENOME", PAGES 1434-1435, SEE ENTIRE DOCUMENT.	1-18



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:	*T	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
A document defining the general state of the art which is not considered to be of particular relevance	*X*	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
E earlier document published on or after the international filing date	*Y*	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
L document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	*G*	document member of the same patent family
O document referring to an oral disclosure, use, exhibition or other means		
P document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search

03 APRIL 1995

Date of mailing of the international search report

17 MAY 1995

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer

EGGERTON CAMPBELL

Telephone No. (703) 308-0196

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US95/01863

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	SCIENCE, VOLUME 252, ISSUED 21 JUNE 1991, ADAMS ET AL, "COMPLEMENTARY DNA SEQUENCING: EXPRESSED SEQUENCE TAGS AND HUMAN GENOME PROJECT", PAGES 1651-1656, SEE ENTIRE DOCUMENT.	1-18

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁶ : C12Q 1/68, G06F 15/00	A1	(11) International Publication Number: WO 95/20681 (43) International Publication Date: 3 August 1995 (03.08.95)
(21) International Application Number: PCT/US95/01160 (22) International Filing Date: 27 January 1995 (27.01.95) (30) Priority Data: 08/187,530 27 January 1994 (27.01.94) US 08/282,955 29 July 1994 (29.07.94) US (71) Applicant: INCYTE PHARMACEUTICALS, INC. [US/US]; 3330 Hillview Avenue, Palo Alto, CA 94304 (US). (72) Inventors: SEILHAMER, Jeffrey, J.; 12555 La Cresta, Los Altos Hills, CA 94022 (US). SCOTT, Randal, W.; 13140 Sun-Mor, Mountain View, CA 94040 (US). (74) Agents: CAGE, Kenneth, L. et al.; Willian Brinks Hofer Gilson & Lione, 2000 K Street, N.W., Suite 200, Washington, DC 20006-1809 (US).		(81) Designated States: AM, AU, BB, BG, BR, BY, CA, CN, CZ, EE, FI, GE, HU, JP, KG, KP, KR, KZ, LK, LR, LT, LV, MD, MG, MN, MX, NO, NZ, PL, RO, RU, SI, SK, TJ, TT, UA, UZ, VN, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG), ARIPO patent (KE, MW, SD, SZ). Published <i>With international search report.</i>
(54) Title: COMPARATIVE GENE TRANSCRIPT ANALYSIS (57) Abstract A method and system for quantifying the relative abundance of gene transcripts in a biological specimen. One embodiment of the method generates high-throughput sequence-specific analysis of multiple RNAs or their corresponding cDNAs (gene transcript imaging analysis). Another embodiment of the method produces a gene transcript imaging analysis by the use of high-throughput cDNA sequence analysis. In addition, the gene transcript imaging can be used to detect or diagnose a particular biological state, disease, or condition which is correlated to the relative abundance of gene transcripts in a given cell or population of cells. The invention provides a method for comparing the gene transcript image analysis from two or more different biological specimens in order to distinguish between the two specimens and identify one or more genes which are differentially expressed between the two specimens.		

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	GB	United Kingdom	MR	Mauritania
AU	Australia	GE	Georgia	MW	Malawi
BB	Barbados	GN	Guinea	NE	Niger
BE	Belgium	GR	Greece	NL	Netherlands
BF	Burkina Faso	HU	Hungary	NO	Norway
BG	Bulgaria	IE	Ireland	NZ	New Zealand
BJ	Benin	IT	Italy	PL	Poland
BR	Brazil	JP	Japan	PT	Portugal
BY	Belarus	KE	Kenya	RO	Romania
CA	Canada	KG	Kyrgystan	RU	Russian Federation
CF	Central African Republic	KP	Democratic People's Republic of Korea	SD	Sudan
CG	Congo	KR	Republic of Korea	SE	Sweden
CH	Switzerland	KZ	Kazakhstan	SI	Slovenia
CI	Côte d'Ivoire	LI	Liechtenstein	SK	Slovakia
CM	Cameroon	LK	Sri Lanka	SN	Senegal
CN	China	LU	Luxembourg	TD	Chad
CS	Czechoslovakia	LV	Latvia	TG	Togo
CZ	Czech Republic	MC	Monaco	TJ	Tajikistan
DE	Germany	MD	Republic of Moldova	TT	Trinidad and Tobago
DK	Denmark	MG	Madagascar	UA	Ukraine
ES	Spain	ML	Mali	US	United States of America
FI	Finland	MN	Mongolia	UZ	Uzbekistan
FR	France			VN	Viet Nam
GA	Gabon				

COMPARATIVE GENE TRANSCRIPT ANALYSIS

1. FIELD OF INVENTION

The present invention is in the field of molecular biology and computer science; more particularly, the present invention describes methods of analyzing gene transcripts and diagnosing the genetic expression of cells and tissue.

2. BACKGROUND OF THE INVENTION

Until very recently, the history of molecular biology has been written one gene at a time. Scientists have observed the cell's physical changes, isolated mixtures from the cell or its milieu, purified proteins, sequenced proteins and therefrom constructed probes to look for the corresponding gene.

Recently, different nations have set up massive projects to sequence the billions of bases in the human genome. These projects typically begin with dividing the genome into large portions of chromosomes and then determining the sequences of these pieces, which are then analyzed for identity with known proteins or portions thereof, known as motifs. Unfortunately, the majority of genomic DNA does not encode proteins and though it is postulated to have some effect on the cell's ability to make protein, its relevance to medical applications is not understood at this time.

A third methodology involves sequencing only the transcripts encoding the cellular machinery actively involved in making protein, namely the mRNA. The advantage is that the cell has already edited out all the non-coding DNA, and it is relatively easy to identify the protein-coding portion of the RNA. The utility of this approach was not immediately obvious to genomic researchers. In fact, when cDNA sequencing was initially proposed, the method was roundly denounced by those committed to genomic sequencing. For example, the head of the U.S. Human Genome project discounted cDNA sequencing as not valuable and refused to approve funding of projects.

In this disclosure, we teach methods for analyzing DNA, including cDNA libraries. Based on our analyses and

research, we see each individual gene product as a "pixel" of information, which relates to the expression of that, and only that, gene. We teach herein, methods whereby the individual "pixels" of gene expression information can be
5 combined into a single gene transcript "image," in which each of the individual genes can be visualized simultaneously and allowing relationships between the gene pixels to be easily visualized and understood.

We further teach a new method which we call electronic
10 subtraction. Electronic subtraction will enable the gene researcher to turn a single image into a moving picture, one which describes the temporality or dynamics of gene expression, at the level of a cell or a whole tissue. It is that sense of "motion" of cellular machinery on the
15 scale of a cell or organ which constitutes the new invention herein. This constitutes a new view into the process of living cell physiology and one which holds great promise to unveil and discover new therapeutic and diagnostic approaches in medicine.

20 We teach another method which we call "electronic northern," which tracks the expression of a single gene across many types of cells and tissues.

Nucleic acids (DNA and RNA) carry within their sequence the hereditary information and are therefore the
25 prime molecules of life. Nucleic acids are found in all living organisms including bacteria, fungi, viruses, plants and animals. It is of interest to determine the relative abundance of different discrete nucleic acids in different cells, tissues and organisms over time under various
30 conditions, treatments and regimes.

All dividing cells in the human body contain the same set of 23 pairs of chromosomes. It is estimated that these autosomal and sex chromosomes encode approximately 100,000 genes. The differences among different types of cells are
35 believed to reflect the differential expression of the 100,000 or so genes. Fundamental questions of biology could be answered by understanding which genes are transcribed and knowing the relative abundance of transcripts in different cells.

Previously, the art has only provided for the analysis of a few known genes at a time by standard molecular biology techniques such as PCR, northern blot analysis, or other types of DNA probe analysis such as in situ hybridization. Each of these methods allows one to analyze the transcription of only known genes and/or small numbers of genes at a time. Nucl. Acids Res. 19, 7097-7104 (1991); Nucl. Acids Res. 18, 4833-42 (1990); Nucl. Acids Res. 18, 2789-92 (1989); European J. Neuroscience 2, 1063-1073 (1990); Analytical Biochem. 187, 364-73 (1990); Genet. Annals Techn. Appl. 7, 64-70 (1990); GATA 8(4), 129-33 (1991); Proc. Natl. Acad. Sci. USA 85, 1696-1700 (1988); Nucl. Acids Res. 19, 1954 (1991); Proc. Natl. Acad. Sci. USA 88, 1943-47 (1991); Nucl. Acids Res. 19, 6123-27 (1991); Proc. Natl. Acad. Sci. USA 85, 5738-42 (1988); Nucl. Acids Res. 16, 10937 (1988).

Studies of the number and types of genes whose transcription is induced or otherwise regulated during cell processes such as activation, differentiation, aging, viral transformation, morphogenesis, and mitosis have been pursued for many years, using a variety of methodologies. One of the earliest methods was to isolate and analyze levels of the proteins in a cell, tissue, organ system, or even organisms both before and after the process of interest. One method of analyzing multiple proteins in a sample is using 2-dimensional gel electrophoresis, wherein proteins can be, in principle, identified and quantified as individual bands, and ultimately reduced to a discrete signal. At present, 2-dimensional analysis only resolves approximately 15% of the proteins. In order to positively analyze those bands which are resolved, each band must be excised from the membrane and subjected to protein sequence analysis using Edman degradation. Unfortunately, most of the bands were present in quantities too small to obtain a reliable sequence, and many of those bands contained more than one discrete protein. An additional difficulty is that many of the proteins were blocked at the amino-terminus, further complicating the sequencing process.

Analyzing differentiation at the gene transcription level has overcome many of these disadvantages and drawbacks, since the power of recombinant DNA technology allows amplification of signals containing very small amounts of material. The most common method, called "hybridization subtraction," involves isolation of mRNA from the biological specimen before (B) and after (A) the developmental process of interest, transcribing one set of mRNA into cDNA, subtracting specimen B from specimen A (mRNA from cDNA) by hybridization, and constructing a cDNA library from the non-hybridizing mRNA fraction. Many different groups have used this strategy successfully, and a variety of procedures have been published and improved upon using this same basic scheme. Nucl. Acids Res. 19, 7097-7104 (1991); Nucl. Acids Res. 18, 4833-42 (1990); Nucl. Acids Res. 18, 2789-92 (1989); European J. Neuroscience 2, 1063-1073 (1990); Analytical Biochem. 187, 364-73 (1990); Genet. Annals Techn. Appl. 7, 64-70 (1990); GATA 8(4), 129-33 (1991); Proc. Natl. Acad. Sci. USA 85, 1696-1700 (1988); Nucl. Acids Res. 19, 1954 (1991); Proc. Natl. Acad. Sci. USA 88, 1943-47 (1991); Nucl. Acids Res. 19, 6123-27 (1991); Proc. Natl. Acad. Sci. USA 85, 5738-42 (1988); Nucl. Acids Res. 16, 10937 (1988).

Although each of these techniques have particular strengths and weaknesses, there are still some limitations and undesirable aspects of these methods: First, the time and effort required to construct such libraries is quite large. Typically, a trained molecular biologist might expect construction and characterization of such a library to require 3 to 6 months, depending on the level of skill, experience, and luck. Second, the resulting subtraction libraries are typically inferior to the libraries constructed by standard methodology. A typical conventional cDNA library should have a clone complexity of at least 10^6 clones, and an average insert size of 1-3 kB. In contrast, subtracted libraries can have complexities of 10^2 or 10^3 and average insert sizes of 0.2 kB. Therefore, there can be a significant loss of clone and sequence information associated with such libraries. Third, this

approach allows the researcher to capture only the genes induced in specimen A relative to specimen B, not vice-versa, nor does it easily allow comparison to a third specimen of interest (C). Fourth, this approach requires very large amounts (hundreds of micrograms) of "driver" mRNA (specimen B), which significantly limits the number and type of subtractions that are possible since many tissues and cells are very difficult to obtain in large quantities.

Fifth, the resolution of the subtraction is dependent upon the physical properties of DNA:DNA or RNA:DNA hybridization. The ability of a given sequence to find a hybridization match is dependent on its unique CoT value. The CoT value is a function of the number of copies (concentration) of the particular sequence, multiplied by the time of hybridization. It follows that for sequences which are abundant, hybridization events will occur very rapidly (low CoT value), while rare sequences will form duplexes at very high CoT values. CoT values which allow such rare sequences to form duplexes and therefore be effectively selected are difficult to achieve in a convenient time frame. Therefore, hybridization subtraction is simply not a useful technique with which to study relative levels of rare mRNA species. Sixth, this problem is further complicated by the fact that duplex formation is also dependent on the nucleotide base composition for a given sequence. Those sequences rich in G + C form stronger duplexes than those with high contents of A + T. Therefore, the former sequences will tend to be removed selectively by hybridization subtraction. Seventh, it is possible that hybridization between nonexact matches can occur. When this happens, the expression of a homologous gene may "mask" expression of a gene of interest, artificially skewing the results for that particular gene.

Matsubara and Okubo proposed using partial cDNA sequences to establish expression profiles of genes which could be used in functional analyses of the human genome. Matsubara and Okubo warned against using random priming, as

it creates multiple unique DNA fragments from individual mRNAs and may thus skew the analysis of the number of particular mRNAs per library. They sequenced randomly selected members from a 3'-directed cDNA library and
5 established the frequency of appearance of the various ESTs. They proposed comparing lists of ESTs from various cell types to classify genes. Genes expressed in many different cell types were labeled housekeepers and those selectively expressed in certain cells were labeled cell-
10 specific genes, even in the absence of the full sequence of the gene or the biological activity of the gene product.

The present invention avoids the drawbacks of the prior art by providing a method to quantify the relative abundance of multiple gene transcripts in a given
15 biological specimen by the use of high-throughput sequence-specific analysis of individual RNAs and/or their corresponding cDNAs.

The present invention offers several advantages over current protein discovery methods which attempt to isolate
20 individual proteins based upon biological effects. The method of the instant invention provides for detailed diagnostic comparisons of cell profiles revealing numerous changes in the expression of individual transcripts.

The instant invention provides several advantages over
25 current subtraction methods including a more complex library analysis (10^6 to 10^7 clones as compared to 10^3 clones) which allows identification of low abundance messages as well as enabling the identification of messages which either increase or decrease in abundance. These
30 large libraries are very routine to make in contrast to the libraries of previous methods. In addition, homologues can easily be distinguished with the method of the instant invention.

This method is very convenient because it organizes a
35 large quantity of data into a comprehensible, digestible format. The most significant differences are highlighted by electronic subtraction. In depth analyses are made more convenient.

The present invention provides several advantages over previous methods of electronic analysis of cDNA. The method is particularly powerful when more than 100 and preferably more than 1,000 gene transcripts are analyzed.

5 In such a case, new low-frequency transcripts are discovered and tissue typed.

High resolution analysis of gene expression can be used directly as a diagnostic profile or to identify disease-specific genes for the development of more classic
10 diagnostic approaches.

This process is defined as gene transcript frequency analysis. The resulting quantitative analysis of the gene transcripts is defined as comparative gene transcript analysis.

15 3. SUMMARY OF THE INVENTION

The invention is a method of analyzing a specimen containing gene transcripts comprising the steps of (a) producing a library of biological sequences; (b) generating a set of transcript sequences, where each of the transcript
20 sequences in said set is indicative of a different one of the biological sequences of the library; (c) processing the transcript sequences in a programmed computer (in which a database of reference transcript sequences indicative of reference sequences is stored), to generate an identified
25 sequence value for each of the transcript sequences, where each said identified sequence value is indicative of sequence annotation and a degree of match between one of the biological sequences of the library and at least one of the reference sequences; and (d) processing each said
30 identified sequence value to generate final data values indicative of the number of times each identified sequence value is present in the library.

The invention also includes a method of comparing two specimens containing gene transcripts. The first specimen
35 is processed as described above. The second specimen is used to produce a second library of biological sequences, which is used to generate a second set of transcript sequences, where each of the transcript sequences in the

In a further embodiment, the relative abundance of the gene transcripts in one cell type or tissue is compared with the relative abundance of gene transcript numbers in a second cell type or tissue in order to identify the
5 differences and similarities.

In a further embodiment, the method includes a system for analyzing a library of biological sequences including a means for receiving a set of transcript sequences, where each of the transcript sequences is indicative of a
10 different one of the biological sequences of the library; and a means for processing the transcript sequences in a computer system in which a database of reference transcript sequences indicative of reference sequences is stored, wherein the computer is programmed with software for
15 generating an identified sequence value for each of the transcript sequences, where each said identified sequence value is indicative of a sequence annotation and the degree of match between a different one of the biological sequences of the library and at least one of the reference
20 sequences, and for processing each said identified sequence value to generate final data values indicative of the number of times each identified sequence value is present in the library.

In essence, the invention is a method and system for
25 quantifying the relative abundance of gene transcripts in a biological specimen. The invention provides a method for comparing the gene transcript image from two or more different biological specimens in order to distinguish between the two specimens and identify one or more genes
30 which are differentially expressed between the two specimens. Thus, this gene transcript image and its comparison can be used as a diagnostic. One embodiment of the method generates high-throughput sequence-specific analysis of multiple RNAs or their corresponding cDNAs: a
35 gene transcript image. Another embodiment of the method produces the gene transcript imaging analysis by the use of high-throughput cDNA sequence analysis. In addition, two or more gene transcript images can be compared and used to detect or diagnose a particular biological state, disease,

or condition which is correlated to the relative abundance of gene transcripts in a given cell or population of cells.

4. DESCRIPTION OF THE TABLES AND DRAWINGS

4.1. TABLES

5 Table 1 presents a detailed explanation of the letter codes utilized in Tables 2-5.

Table 2 lists the one hundred most common gene transcripts. It is a partial list of isolates from the HUVEC cDNA library prepared and sequenced as described
10 below. The left-hand column refers to the sequence's order of abundance in this table. The next column labeled "number" is the clone number of the first HUVEC sequence identification reference matching the sequence in the "entry" column number. Isolates that have not been
15 sequenced are not present in Table 2. The next column, labeled "N", indicates the total number of cDNAs which have the same degree of match with the sequence of the reference transcript in the "entry" column.

 The column labeled "entry" gives the NIH GENBANK locus
20 name, which corresponds to the library sequence numbers. The "s" column indicates in a few cases the species of the reference sequence. The code for column "s" is given in Table 1. The column labeled "descriptor" provides a plain English explanation of the identity of the sequence
25 corresponding to the NIH GENBANK locus name in the "entry" column.

Table 3 is a comparison of the top fifteen most abundant gene transcripts in normal monocytes and activated macrophage cells.

30 Table 4 is a detailed summary of library subtraction analysis summary comparing the THP-1 and human macrophage cDNA sequences. In Table 4, the same code as in Table 2 is used. Additional columns are for "bgfreq" (abundance number in the subtractant library), "rfend" (abundance
35 number in the target library) and "ratio" (the target abundance number divided by the subtractant abundance number). As is clear from perusal of the table, when the abundance number in the subtractant library is "0", the

target abundance number is divided by 0.05. This is a way of obtaining a result (not possible dividing by 0) and distinguishing the result from ratios of subtractant numbers of 1.

5 Table 5 is the computer program, written in source code, for generating gene transcript subtraction profiles.

Table 6 is a partial listing of database entries used in the electronic northern blot analysis as provided by the present invention.

10

4.2. BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a chart summarizing data collected and stored regarding the library construction portion of sequence preparation and analysis.

15 Figure 2 is a diagram representing the sequence of operations performed by "abundance sort" software in a class of preferred embodiments of the inventive method.

Figure 3 is a block diagram of a preferred embodiment of the system of the invention.

20 Figure 4 is a more detailed block diagram of the bioinformatics process from new sequence (that has already been sequenced but not identified) to printout of the transcript imaging analysis and the provision of database subscriptions.

25 5. DETAILED DESCRIPTION OF THE INVENTION

The present invention provides a method to compare the relative abundance of gene transcripts in different biological specimens by the use of high-throughput sequence-specific analysis of individual RNAs or their
30 corresponding cDNAs (or alternatively, of data representing other biological sequences). This process is denoted herein as gene transcript imaging. The quantitative analysis of the relative abundance for a set of gene transcripts is denoted herein as "gene transcript image
35 analysis" or "gene transcript frequency analysis". The present invention allows one to obtain a profile for gene transcription in any given population of cells or tissue from any type of organism. The invention can be applied to

obtain a profile of a specimen consisting of a single cell (or clones of a single cell), or of many cells, or of tissue more complex than a single cell and containing multiple cell types, such as liver.

- 5 The invention has significant advantages in the fields of diagnostics, toxicology and pharmacology, to name a few. A highly sophisticated diagnostic test can be performed on the ill patient in whom a diagnosis has not been made. A biological specimen consisting of the patient's fluids or
10 tissues is obtained, and the gene transcripts are isolated and expanded to the extent necessary to determine their identity. Optionally, the gene transcripts can be converted to cDNA. A sampling of the gene transcripts are subjected to sequence-specific analysis and quantified.
- 15 These gene transcript sequence abundances are compared against reference database sequence abundances including normal data sets for diseased and healthy patients. The patient has the disease(s) with which the patient's data set most closely correlates.

- 20 For example, gene transcript frequency analysis can be used to differentiate normal cells or tissues from diseased cells or tissues, just as it highlights differences between normal monocytes and activated macrophages in Table 3.

- In toxicology, a fundamental question is which tests
25 are most effective in predicting or detecting a toxic effect. Gene transcript imaging provides highly detailed information on the cell and tissue environment, some of which would not be obvious in conventional, less detailed screening methods. The gene transcript image is a more
30 powerful method to predict drug toxicity and efficacy. Similar benefits accrue in the use of this tool in pharmacology. The gene transcript image can be used selectively to look at protein categories which are expected to be affected, for example, enzymes which
35 detoxify toxins.

 In an alternative embodiment, comparative gene transcript frequency analysis is used to differentiate between cancer cells which respond to anti-cancer agents and those which do not respond. Examples of anti-cancer

agents are tamoxifen, vincristine, vinblastine, podophyllotoxins, etoposide, teniposide, cisplatin, biologic response modifiers such as interferon, Il-2, GM-CSF, enzymes, hormones and the like. This method also provides a means for sorting the gene transcripts by functional category. In the case of cancer cells, transcription factors or other essential regulatory molecules are very important categories to analyze across different libraries.

10 In yet another embodiment, comparative gene transcript frequency analysis is used to differentiate between control liver cells and liver cells isolated from patients treated with experimental drugs like FIAU to distinguish between pathology caused by the underlying disease and that caused by the drug.

In yet another embodiment, comparative gene transcript frequency analysis is used to differentiate between brain tissue from patients treated and untreated with lithium.

20 In a further embodiment, comparative gene transcript frequency analysis is used to differentiate between cyclosporin and FK506-treated cells and normal cells.

In a further embodiment, comparative gene transcript frequency analysis is used to differentiate between virally infected (including HIV-infected) human cells and uninfected human cells. Gene transcript frequency analysis is also used to rapidly survey gene transcripts in HIV-resistant, HIV-infected, and HIV-sensitive cells. Comparison of gene transcript abundance will indicate the success of treatment and/or new avenues to study.

30 In a further embodiment, comparative gene transcript frequency analysis is used to differentiate between bronchial lavage fluids from healthy and unhealthy patients with a variety of ailments.

In a further embodiment, comparative gene transcript frequency analysis is used to differentiate between cell, plant, microbial and animal mutants and wild-type species. In addition, the transcript abundance program is adapted to permit the scientist to evaluate the transcription of one gene in many different tissues. Such comparisons could

identify deletion mutants which do not produce a gene product and point mutants which produce a less abundant or otherwise different message. Such mutations can affect basic biochemical and pharmacological processes, such as mineral nutrition and metabolism, and can be isolated by means known to those skilled in the art. Thus, crops with improved yields, pest resistance and other factors can be developed.

In a further embodiment, comparative gene transcript frequency analysis is used for an interspecies comparative analysis which would allow for the selection of better pharmacologic animal models. In this embodiment, humans and other animals (such as a mouse), or their cultured cells are treated with a specific test agent. The relative sequence abundance of each cDNA population is determined. If the animal test system is a good model, homologous genes in the animal cDNA population should change expression similarly to those in human cells. If side effects are detected with the drug, a detailed transcript abundance analysis will be performed to survey gene transcript changes. Models will then be evaluated by comparing basic physiological changes.

In a further embodiment, comparative gene transcript frequency analysis is used in a clinical setting to give a highly detailed gene transcript profile of a patient's cells or tissue (for example, a blood sample). In particular, gene transcript frequency analysis is used to give a high resolution gene expression profile of a diseased state or condition.

In the preferred embodiment, the method utilizes high-throughput cDNA sequencing to identify specific transcripts of interest. The generated cDNA and deduced amino acid sequences are then extensively compared with GENBANK and other sequence data banks as described below. The method offers several advantages over current protein discovery by two-dimensional gel methods which try to identify individual proteins involved in a particular biological effect. Here, detailed comparisons of profiles of activated and inactive cells reveal numerous changes in

the expression of individual transcripts. After it is determined if the sequence is an "exact" match, similar or a non-match, the sequence is entered into a database. Next, the numbers of copies of cDNA corresponding to each gene are tabulated. Although this can be done slowly and arduously, if at all, by human hand from a printout of all entries, a computer program is a useful and rapid way to tabulate this information. The numbers of cDNA copies (optionally divided by the total number of sequences in the data set) provides a picture of the relative abundance of transcripts for each corresponding gene. The list of represented genes can then be sorted by abundance in the cDNA population. A multitude of additional types of comparisons or dimensions are possible and are exemplified below.

An alternate method of producing a gene transcript image includes the steps of obtaining a mixture of test mRNA and providing a representative array of unique probes whose sequences are complementary to at least some of the test mRNAs. Next, a fixed amount of the test mRNA is added to the arrayed probes. The test mRNA is incubated with the probes for a sufficient time to allow hybrids of the test mRNA and probes to form. The mRNA-probe hybrids are detected and the quantity determined. The hybrids are identified by their location in the probe array. The quantity of each hybrid is summed to give a population number. Each hybrid quantity is divided by the population number to provide a set of relative abundance data termed a gene transcript image analysis.

30

6. EXAMPLES

The examples below are provided to illustrate the subject invention. These examples are provided by way of illustration and are not included for the purpose of limiting the invention.

35

6.1. TISSUE SOURCES AND CELL LINES

For analysis with the computer program claimed herein, biological sequences can be obtained from virtually any

source. Most popular are tissues obtained from the human body. Tissues can be obtained from any organ of the body, any age donor, any abnormality or any immortalized cell line. Immortal cell lines may be preferred in some instances because of their purity of cell type; other tissue samples invariably include mixed cell types. A special technique is available to take a single cell (for example, a brain cell) and harness the cellular machinery to grow up sufficient cDNA for sequencing by the techniques and analysis described herein (cf. U.S. Patent Nos. 5,021,335 and 5,168,038, which are incorporated by reference). The examples given herein utilized the following immortalized cell lines: monocyte-like U-937 cells, activated macrophage-like THP-1 cells, induced vascular endothelial cells (HUVEC cells) and mast cell-like HMC-1 cells.

The U-937 cell line is a human histiocytic lymphoma cell line with monocyte characteristics, established from malignant cells obtained from the pleural effusion of a patient with diffuse histiocytic lymphoma (Sundstrom, C. and Nilsson, K. (1976) Int. J. Cancer 17:565). U-937 is one of only a few human cell lines with the morphology, cytochemistry, surface receptors and monocyte-like characteristics of histiocytic cells. These cells can be induced to terminal monocytic differentiation and will express new cell surface molecules when activated with supernatants from human mixed lymphocyte cultures. Upon this type of in vitro activation, the cells undergo morphological and functional changes, including augmentation of antibody-dependent cellular cytotoxicity (ADCC) against erythroid and tumor target cells (one of the principal functions of macrophages). Activation of U-937 cells with phorbol 12-myristate 13-acetate (PMA) in vitro stimulates the production of several compounds, including prostaglandins, leukotrienes and platelet-activating factor (PAF), which are potent inflammatory mediators. Thus, U-937 is a cell line that is well suited for the identification and isolation of gene transcripts associated with normal monocytes.

The HUVEC cell line is a normal, homogeneous, well characterized, early passage endothelial cell culture from human umbilical vein (Cell Systems Corp., 12815 NE 124th Street, Kirkland, WA 98034). Only gene transcripts from
5 induced, or treated, HUVEC cells were sequenced. One batch of 1×10^8 cells was treated for 5 hours with 1 U/ml rIL-1b and 100 ng/ml E.coli lipopolysaccharide (LPS) endotoxin prior to harvesting. A separate batch of 2×10^8 cells was treated at confluence with 4 U/ml TNF and 2 U/ml
10 interferon-gamma (IFN-gamma) prior to harvesting.

THP-1 is a human leukemic cell line with distinct monocytic characteristics. This cell line was derived from the blood of a 1-year-old boy with acute monocytic leukemia (Tsuchiya, S. et al. (1980) Int. J. Cancer: 171-76). The
15 following cytological and cytochemical criteria were used to determine the monocytic nature of the cell line: 1) the presence of alpha-naphthyl butyrate esterase activity which could be inhibited by sodium fluoride; 2) the production of lysozyme; 3) the phagocytosis of latex particles and
20 sensitized SRBC (sheep red blood cells); and 4) the ability of mitomycin C-treated THP-1 cells to activate T-lymphocytes following ConA (concanavalin A) treatment. Morphologically, the cytoplasm contained small azurophilic granules and the nucleus was indented and irregularly
25 shaped with deep folds. The cell line had Fc and C3b receptors, probably functioning in phagocytosis. THP-1 cells treated with the tumor promoter 12-o-tetradecanoyl-phorbol-13 acetate (TPA) stop proliferating and differentiate into macrophage-like cells which mimic native
30 monocyte-derived macrophages in several respects. Morphologically, as the cells change shape, the nucleus becomes more irregular and additional phagocytic vacuoles appear in the cytoplasm. The differentiated THP-1 cells also exhibit an increased adherence to tissue culture
35 plastic.

HMC-1 cells (a human mast cell line) were established from the peripheral blood of a Mayo Clinic patient with mast cell leukemia (Leukemia Res. (1988) 12:345-55). The cultured cells looked similar to immature cloned murine

mast cells, contained histamine, and stained positively for chloroacetate esterase, amino caproate esterase, eosinophil major basic protein (MBP) and tryptase. The HMC-1 cells have, however, lost the ability to synthesize normal IgE
5 receptors. HMC-1 cells also possess a 10;16 translocation, present in cells initially collected by leukophoresis from the patient and not an artifact of culturing. Thus, HMC-1 cells are a good model for mast cells.

6.2. CONSTRUCTION OF cDNA LIBRARIES

10 For inter-library comparisons, the libraries must be prepared in similar manners. Certain parameters appear to be particularly important to control. One such parameter is the method of isolating mRNA. It is important to use the same conditions to remove DNA and heterogeneous nuclear
15 RNA from comparison libraries. Size fractionation of cDNA must be carefully controlled. The same vector preferably should be used for preparing libraries to be compared. At the very least, the same type of vector (e.g., unidirectional vector) should be used to assure a valid
20 comparison. A unidirectional vector may be preferred in order to more easily analyze the output.

It is preferred to prime only with oligo dT unidirectional primer in order to obtain one only clone per mRNA transcript when obtaining cDNAs. However, it is
25 recognized that employing a mixture of oligo dT and random primers can also be advantageous because such a mixture results in more sequence diversity when gene discovery also is a goal. Similar effects can be obtained with DR2 (Clontech) and HXLOX (US Biochemical) and also vectors from
30 Invitrogen and Novagen. These vectors have two requirements. First, there must be primer sites for commercially available primers such as T3 or M13 reverse primers. Second, the vector must accept inserts up to 10 kB.

35 It also is important that the clones be randomly sampled, and that a significant population of clones is used. Data have been generated with 5,000 clones; however, if very rare genes are to be obtained and/or their relative

abundance determined, as many as 100,000 clones from a single library may need to be sampled. Size fractionation of cDNA also must be carefully controlled. Alternately, plaques can be selected, rather than clones.

5 Besides the Uni-ZAP™ vector system by Stratagene disclosed below, it is now believed that other similarly unidirectional vectors also can be used. For example, it is believed that such vectors include but are not limited to DR2 (Clontech), and HXLOX (U.S. Biochemical).

10 Preferably, the details of library construction (as shown in Figure 1) are collected and stored in a database for later retrieval relative to the sequences being compared. Fig. 1 shows important information regarding the library collaborator or cell or cDNA supplier,
15 pretreatment, biological source, culture, mRNA preparation and cDNA construction. Similarly detailed information about the other steps is beneficial in analyzing sequences and libraries in depth.

RNA must be harvested from cells and tissue samples
20 and cDNA libraries are subsequently constructed. cDNA libraries can be constructed according to techniques known in the art. (See, for example, Maniatis, T. et al. (1982) Molecular Cloning, Cold Spring Harbor Laboratory, New York). cDNA libraries may also be purchased. The U-937
25 cDNA library (catalog No. 937207) was obtained from Stratagene, Inc., 11099 M. Torrey Pines Rd., La Jolla, CA 92037.

The THP-1 cDNA library was custom constructed by Stratagene from THP-1 cells cultured 48 hours with 100 nm
30 TPA and 4 hours with 1 µg/ml LPS. The human mast cell HMC-1 cDNA library was also custom constructed by Stratagene from cultured HMC-1 cells. The HUVEC cDNA library was custom constructed by Stratagene from two batches of induced HUVEC cells which were separately processed.

35 Essentially, all the libraries were prepared in the same manner. First, poly(A+)RNA (mRNA) was purified. For the U-937 and HMC-1 RNA, cDNA synthesis was only primed with oligo dT. For the THP-1 and HUVEC RNA, cDNA synthesis was primed separately with both oligo dT and random

hexamers, and the two cDNA libraries were treated separately. Synthetic adaptor oligonucleotides were ligated onto cDNA ends enabling its insertion into the Uni-Zap™ vector system (Stratagene), allowing high efficiency
5 unidirectional (sense orientation) lambda library construction and the convenience of a plasmid system with blue-white color selection to detect clones with cDNA insertions. Finally, the two libraries were combined into a single library by mixing equal numbers of bacteriophage.
10 The libraries can be screened with either DNA probes or antibody probes and the pBluescript® phagemid (Stratagene) can be rapidly excised in vivo. The phagemid allows the use of a plasmid system for easy insert characterization, sequencing, site-directed mutagenesis,
15 the creation of unidirectional deletions and expression of fusion proteins. The custom-constructed library phage particles were infected into E. coli host strain XL1-Blue® (Stratagene), which has a high transformation efficiency, increasing the probability of obtaining rare, under-
20 represented clones in the cDNA library.

6.3. ISOLATION OF cDNA CLONES

The phagemid forms of individual cDNA clones were obtained by the in vivo excision process, in which the host bacterial strain was coinfectd with both the lambda
25 library phage and an f1 helper phage. Proteins derived from both the library-containing phage and the helper phage nicked the lambda DNA, initiated new DNA synthesis from defined sequences on the lambda target DNA and created a smaller, single stranded circular phagemid DNA molecule
30 that included all DNA sequences of the pBluescript® plasmid and the cDNA insert. The phagemid DNA was secreted from the cells and purified, then used to re-infect fresh host cells, where the double stranded phagemid DNA was produced. Because the phagemid carries the gene for beta-lactamase,
35 the newly-transformed bacteria are selected on medium containing ampicillin.

Phagemid DNA was purified using the Magic Minipreps™ DNA Purification System (Promega catalogue #A7100. Promega

Corp., 2800 Woods Hollow Rd., Madison, WI 53711). This small-scale process provides a simple and reliable method for lysing the bacterial cells and rapidly isolating purified phagemid DNA using a proprietary DNA-binding resin. The DNA was eluted from the purification resin already prepared for DNA sequencing and other analytical manipulations.

Phagemid DNA was also purified using the QIAwell-8 Plasmid Purification System from QIAGEN® DNA Purification System (QIAGEN Inc., 9259 Eton Ave., Chatsworth, CA 91311). This product line provides a convenient, rapid and reliable high-throughput method for lysing the bacterial cells and isolating highly purified phagemid DNA using QIAGEN anion-exchange resin particles with EMPORE™ membrane technology from 3M in a multiwell format. The DNA was eluted from the purification resin already prepared for DNA sequencing and other analytical manipulations.

An alternate method of purifying phagemid has recently become available. It utilizes the Miniprep Kit (Catalog No. 77468, available from Advanced Genetic Technologies Corp., 19212 Orbit Drive, Gaithersburg, Maryland). This kit is in the 96-well format and provides enough reagents for 960 purifications. Each kit is provided with a recommended protocol, which has been employed except for the following changes. First, the 96 wells are each filled with only 1 ml of sterile terrific broth with carbenicillin at 25 mg/L and glycerol at 0.4%. After the wells are inoculated, the bacteria are cultured for 24 hours and lysed with 60 µl of lysis buffer. A centrifugation step (2900 rpm for 5 minutes) is performed before the contents of the block are added to the primary filter plate. The optional step of adding isopropanol to TRIS buffer is not routinely performed. After the last step in the protocol, samples are transferred to a Beckman 96-well block for storage.

Another new DNA purification system is the WIZARD™ product line which is available from Promega (catalog No. A7071) and may be adaptable to the 96-well format.

6.4. SEQUENCING OF cDNA CLONES

The cDNA inserts from random isolates of the U-937 and THP-1 libraries were sequenced in part. Methods for DNA sequencing are well known in the art. Conventional enzymatic methods employ DNA polymerase Klenow fragment, Sequenase™ or Taq polymerase to extend DNA chains from an oligonucleotide primer annealed to the DNA template of interest. Methods have been developed for the use of both single- and double-stranded templates. The chain termination reaction products are usually electrophoresed on urea-acrylamide gels and are detected either by autoradiography (for radionuclide-labeled precursors) or by fluorescence (for fluorescent-labeled precursors). Recent improvements in mechanized reaction preparation, sequencing and analysis using the fluorescent detection method have permitted expansion in the number of sequences that can be determined per day (such as the Applied Biosystems 373 and 377 DNA sequencer, Catalyst 800). Currently with the system as described, read lengths range from 250 to 400 bases and are clone dependent. Read length also varies with the length of time the gel is run. In general, the shorter runs tend to truncate the sequence. A minimum of only about 25 to 50 bases is necessary to establish the identification and degree of homology of the sequence. Gene transcript imaging can be used with any sequence-specific method, including, but not limited to hybridization, mass spectroscopy, capillary electrophoresis and 505 gel electrophoresis.

30 6.5. HOMOLOGY SEARCHING OF cDNA CLONE AND DEDUCED PROTEIN (and Subsequent Steps)

Using the nucleotide sequences derived from the cDNA clones as query sequences (sequences of a Sequence Listing), databases containing previously identified sequences are searched for areas of homology (similarity). Examples of such databases include Genbank and EMBL. We next describe examples of two homology search algorithms that can be used, and then describe the subsequent computer-implemented steps to be performed in accordance with preferred embodiments of the invention.

In the following description of the computer-implemented steps of the invention, the word "library" denotes a set (or population) of biological specimen nucleic acid sequences. A "library" can consist of cDNA sequences, RNA sequences, or the like, which characterize a biological specimen. The biological specimen can consist of cells of a single human cell type (or can be any of the other above-mentioned types of specimens). We contemplate that the sequences in a library have been determined so as to accurately represent or characterize a biological specimen (for example, they can consist of representative cDNA sequences from clones of RNA taken from a single human cell).

In the following description of the computer-implemented steps of the invention, the expression "database" denotes a set of stored data which represent a collection of sequences, which in turn represent a collection of biological reference materials. For example, a database can consist of data representing many stored cDNA sequences which are in turn representative of human cells infected with various viruses, cells of humans of various ages, cells from different mammalian species, and so on.

In preferred embodiments, the invention employs a computer programmed with software (to be described) for performing the following steps:

(a) processing data indicative of a library of cDNA sequences (generated as a result of high-throughput cDNA sequencing or other method) to determine whether each sequence in the library matches a DNA sequence of a reference database of DNA sequences (and if so, identifying the reference database entry which matches the sequence and indicating the degree of match between the reference sequence and the library sequence) and assigning an identified sequence value based on the sequence annotation and degree of match to each of the sequences in the library;

(b) for some or all entries of the database, tabulating the number of matching identified sequence

values in the library (Although this can be done by human hand from a printout of all entries, we prefer to perform this step using computer software to be described below.), thereby generating a set of final data values or "abundance numbers"; and

(c) if the libraries are different sizes, dividing each abundance number by the total number of sequences in the library, to obtain a relative abundance number for each identified sequence value (i.e., a relative abundance of each gene transcript).

The list of identified sequence values (or genes corresponding thereto) can then be sorted by abundance in the cDNA population. A multitude of additional types of comparisons or dimensions are possible.

For example (to be described below in greater detail), steps (a) and (b) can be repeated for two different libraries (sometimes referred to as a "target" library and a "subtractant" library). Then, for each identified sequence value (or gene transcript), a "ratio" value is obtained by dividing the abundance number (for that identified sequence value) for the target library, by the abundance number (for that identified sequence value) for the subtractant library.

In fact, subtraction may be carried out on multiple libraries. It is possible to add the transcripts from several libraries (for example, three) and then to divide them by another set of transcripts from multiple libraries (again, for example, three). Notation for this operation may be abbreviated as $(A+B+C) / (D+E+F)$, where the capital letters each indicate an entire library. Optionally the abundance numbers of transcripts in the summed libraries may be divided by the total sample size before subtraction.

Unlike standard hybridization technology which permits a single subtraction of two libraries, once one has processed a set or library transcript sequences and stored them in the computer, any number of subtractions can be performed on the library. For example, by this method, ratio values can be obtained by dividing relative abundance

values in a first library by corresponding values in a second library and vice versa.

In variations on step (a), the library consists of nucleotide sequences derived from cDNA clones. Examples of
5 databases which can be searched for areas of homology (similarity) in step (a) include the commercially available databases known as Genbank (NIH) EMBL (European Molecular Biology Labs, Germany), and GENESEQ (Intelligenetics, Mountain View, California).

10 One homology search algorithm which can be used to implement step (a) is the algorithm described in the paper by D.J. Lipman and W.R. Pearson, entitled "Rapid and Sensitive Protein Similarity Searches," Science, 227:1435 (1985). In this algorithm, the homologous regions are
15 searched in a two-step manner. In the first step, the highest homologous regions are determined by calculating a matching score using a homology score table. The parameter "Ktup" is used in this step to establish the minimum window size to be shifted for comparing two sequences. Ktup also
20 sets the number of bases that must match to extract the highest homologous region among the sequences. In this step, no insertions or deletions are applied and the homology is displayed as an initial (INIT) value.

In the second step, the homologous regions are aligned
25 to obtain the highest matching score by inserting a gap in order to add a probable deleted portion. The matching score obtained in the first step is recalculated using the homology score Table and the insertion score Table to an optimized (OPT) value in the final output.

30 DNA homologies between two sequences can be examined graphically using the Harr method of constructing dot matrix homology plots (Needleman, S.B. and Wunsch, C.O., J. Mom. Biol 48:443 (1970)). This method produces a two-dimensional plot which can be useful in determining
35 regions of homology versus regions of repetition.

However, in a class of preferred embodiments, step (a) is implemented by processing the library data in the commercially available computer program known as the INHERIT 670 Sequence Analysis System, available from

Applied Biosystems Inc. (Foster City, California), including the software known as the Factura software (also available from Applied Biosystems Inc.). The Factura program preprocesses each library sequence to "edit out" portions thereof which are not likely to be of interest, such as the vector used to prepare the library. Additional sequences which can be edited out or masked (ignored by the search tools) include but are not limited to the polyA tail and repetitive GAG and CCC sequences. A low-end search program can be written to mask out such "low-information" sequences, or programs such as BLAST can ignore the low-information sequences.

In the algorithm implemented by the INHERIT 670 Sequence Analysis System, the Pattern Specification Language (developed by TRW Inc.) is used to determine regions of homology. "There are three parameters that determine how INHERIT analysis runs sequence comparisons: window size, window offset and error tolerance. Window size specifies the length of the segments into which the query sequence is subdivided. Window offset specifies where to start the next segment [to be compared], counting from the beginning of the previous segment. Error tolerance specifies the total number of insertions, deletions and/or substitutions that are tolerated over the specified word length. Error tolerance may be set to any integer between 0 and 6. The default settings are window tolerance=20, window offset=10 and error tolerance=3." INHERIT Analysis Users Manual, pp.2-15. Version 1.0, Applied Biosystems, Inc., October 1991.

Using a combination of these three parameters, a database (such as a DNA database) can be searched for sequences containing regions of homology and the appropriate sequences are scored with an initial value. Subsequently, these homologous regions are examined using dot matrix homology plots to determine regions of homology versus regions of repetition. Smith-Waterman alignments can be used to display the results of the homology search. The INHERIT software can be executed by a Sun computer system programmed with the UNIX operating system.

Search alternatives to INHERIT include the BLAST program, GCG (available from the Genetics Computer Group, WI) and the Dasher program (Temple Smith, Boston University, Boston, MA). Nucleotide sequences can be
5 searched against Genbank, EMBL or custom databases such as GENESEQ (available from Intelligenetics, Mountain View, CA) or other databases for genes. In addition, we have searched some sequences against our own in-house database.

In preferred embodiments, the transcript sequences are
10 analyzed by the INHERIT software for best conformance with a reference gene transcript to assign a sequence identifier and assigned the degree of homology, which together are the identified sequence value and are input into, and further processed by, a Macintosh personal computer (available from
15 Apple) programmed with an "abundance sort and subtraction analysis" computer program (to be described below).

Prior to the abundance sort and subtraction analysis program (also denoted as the "abundance sort" program), identified sequences from the cDNA clones are assigned
20 value (according to the parameters given above) by degree of match according to the following categories: "exact" matches (regions with a high degree of identity), homologous human matches (regions of high similarity, but not "exact" matches), homologous non-human matches (regions
25 of high similarity present in species other than human), or non matches (no significant regions of homology to previously identified nucleotide sequences stored in the form of the database). Alternately, the degree of match can be a numeric value as described below.

30 With reference again to the step of identifying matches between reference sequences and database entries, protein and peptide sequences can be deduced from the nucleic acid sequences. Using the deduced polypeptide sequence, the match identification can be performed in a
35 manner analogous to that done with cDNA sequences. A protein sequence is used as a query sequence and compared to the previously identified sequences contained in a database such as the Swiss/Prot, PIR and the NBRF Protein database to find homologous proteins. These proteins are

initially scored for homology using a homology score Table (Orcutt, B.C. and Dayoff, M.O. Scoring Matrices, PIR Report MAT - 0285 (February 1985)) resulting in an INIT score. The homologous regions are aligned to obtain the
5 highest matching scores by inserting a gap which adds a probable deleted portion. The matching score is recalculated using the homology score Table and the insertion score Table resulting in an optimized (OPT) score. Even in the absence of knowledge of the proper
10 reading frame of an isolated sequence, the above-described protein homology search may be performed by searching all 3 reading frames.

Peptide and protein sequence homologies can also be ascertained using the INHERIT 670 Sequence Analysis System
15 in an analogous way to that used in DNA sequence homologies. Pattern Specification Language and parameter windows are used to search protein databases for sequences containing regions of homology which are scored with an initial value. Subsequent display in a dot-matrix homology
20 plot shows regions of homology versus regions of repetition. Additional search tools that are available to use on pattern search databases include PLsearch Blocks (available from Henikoff & Henikoff, University of Washington, Seattle), Dasher and GCG. Pattern search
25 databases include, but are not limited to, Protein Blocks (available from Henikoff & Henikoff, University of Washington, Seattle), Brookhaven Protein (available from the Brookhaven National Laboratory, Brookhaven, MA), PROSITE (available from Amos Bairoch, University of Geneva,
30 Switzerland), ProDom (available from Temple Smith, Boston University), and PROTEIN MOTIF FINGERPRINT (available from University of Leeds, United Kingdom).

The ABI Assembler application software, part of the INHERIT DNA analysis system (available from Applied
35 Biosystems, Inc., Foster City, CA), can be employed to create and manage sequence assembly projects by assembling data from selected sequence fragments into a larger sequence. The Assembler software combines two advanced computer technologies which maximize the ability to

assemble sequenced DNA fragments into Assemblages, a special grouping of data where the relationships between sequences are shown by graphic overlap, alignment and statistical views. The process is based on the

5 Meyers-Kececiloglu model of fragment assembly (INHERIT™ Assembler User's Manual, Applied Biosystems, Inc., Foster City, CA), and uses graph theory as the foundation of a very rigorous multiple sequence alignment engine for assembling DNA sequence fragments. Other assembly programs

10 that can be used include MEGALIGN (available from DNASTAR Inc., Madison, WI), Dasher and STADEN (available from Roger Staden, Cambridge, England).

Next, with reference to Fig. 2, we describe in more detail the "abundance sort" program which implements above-

15 mentioned "step (b)" to tabulate the number of sequences of the library which match each database entry (the "abundance number" for each database entry).

Fig. 2 is a flow chart of a preferred embodiment of the abundance sort program. A source code listing of this

20 embodiment of the abundance sort program is set forth in Table 5. In the Table 5 implementation, the abundance sort program is written using the FoxBASE programming language commercially available from Microsoft Corporation. Although FoxBASE was the program chosen for the first

25 iteration of this technology, it should not be considered limiting. Many other programming languages, Sybase being a particularly desirable alternative, can also be used, as will be obvious to one with ordinary skill in the art. The subroutine names specified in Fig. 2 correspond to

30 subroutines listed in Table 5.

With reference again to Fig. 2, the "Identified Sequences" are transcript sequences representing each sequence of the library and a corresponding identification of the database entry (if any) which it matches. In other

35 words, the "Identified Sequences" are transcript sequences representing the output of above-discussed "step (a)."

Fig. 3 is a block diagram of a system for implementing the invention. The Fig. 3 system includes library generation unit 2 which generates a library and asserts an

output stream of transcript sequences indicative of the biological sequences comprising the library. Programmed processor 4 receives the data stream output from unit 2 and processes this data in accordance with above-discussed

5 "step (a)" to generate the Identified Sequences. Processor 4 can be a processor programmed with the commercially available computer program known as the INHERIT 670 Sequence Analysis System and the commercially available computer program known as the Factura program (both

10 available from Applied Biosystems Inc.) and with the UNIX operating system.

Still with reference to Fig. 3, the Identified Sequences are loaded into processor 6 which is programmed with the abundance sort program. Processor 6 generates the

15 Final Transcript sequences indicated in both Figs. 2 and 3. Fig. 4 shows a more detailed block diagram of a planned relational computer system, including various searching techniques which can be implemented, along with an assortment of databases to query against.

20 With reference to Fig. 2, the abundance sort program first performs an operation known as "Tempnum" on the Identified Sequences, to discard all of the Identified Sequences except those which match database entries of selected types. For example, the Tempnum process can

25 select Identified Sequences which represent matches of the following types with database entries (see above for definition): "exact" matches, human "homologous" matches, "other species" matches representing genes present in species other than human), "no" matches (no significant

30 regions of homology with database entries representing previously identified nucleotide sequences), "I" matches (Incyte for not previously known DNA sequences), or "X" matches (matches ESTs in reference database). This eliminates the U, S, M, V, A, R and D sequence (see Table 1

35 for definitions).

The identified sequence values selected during the "Tempnum" process then undergo a further selection (weeding out) operation known as "Tempred." This operation can, for

example, discard all identified sequence values representing matches with selected database entries.

The identified sequence values selected during the "Tempred" process are then classified according to library, during the "Tempdesig" operation. It is contemplated that the "Identified Sequences" can represent sequences from a single library, or from two or more libraries.

Consider first the case that the identified sequence values represent sequences from a single library. In this case, all the identified sequence values determined during "Tempred" undergo sorting in the "Templib" operation, further sorting in the "Libsort" operation, and finally additional sorting in the "Temptarsort" operation. For example, these three sorting operations can sort the identified sequences in order of decreasing "abundance number" (to generate a list of decreasing abundance numbers, each abundance number corresponding to a unique identified sequence entry, or several lists of decreasing abundance numbers, with the abundance numbers in each list corresponding to database entries of a selected type) with redundancies eliminated from each sorted list. In this case, the operation identified as "Cruncher" can be bypassed, so that the "Final Data" values are the organized transcript sequences produced during the "Temptarsort" operation.

We next consider the case that the transcript sequences produced during the "Tempred" operation represent sequences from two libraries (which we will denote the "target" library and the "subtractant" library). For example, the target library may consist of cDNA sequences from clones of a diseased cell, while the subtractant library may consist of cDNA sequences from clones of the diseased cell after treatment by exposure to a drug. For another example, the target library may consist of cDNA sequences from clones of a cell type from a young human, while the subtractant library may consist of cDNA sequences from clones of the same cell type from the same human at different ages.

In this case, the "Tempdesig" operation routes all transcript sequences representing the target library for processing in accordance with "Templib" (and then "Libsort" and "Temptarsort"), and routes all transcript sequences representing the subtractant library for processing in accordance with "Tempsub" (and then "Subsort" and "Tempsubsort"). For example, the consecutive "Templib," "Libsort," and "Temptarsort" sorting operations sort identified sequences from the target library in order of decreasing abundance number (to generate a list of decreasing abundance numbers, each abundance number corresponding to a database entry, or several lists of decreasing abundance numbers, with the abundance numbers in each list corresponding to database entries of a selected type) with redundancies eliminated from each sorted list. The consecutive "Tempsub," "Subsort," and "Tempsubsort" sorting operations sort identified sequences from the subtractant library in order of decreasing abundance number (to generate a list of decreasing abundance numbers, each abundance number corresponding to a database entry, or several lists of decreasing abundance numbers, with the abundance numbers in each list corresponding to database entries of a selected type) with redundancies eliminated from each sorted list.

The transcript sequences output from the "Temptarsort" operation typically represent sorted lists from which a histogram could be generated in which position along one (e.g., horizontal) axis indicates abundance number (of target library sequences), and position along another (e.g., vertical) axis indicates identified sequence value (e.g., human or non-human gene type). Similarly, the transcript sequences output from the "Tempsubsort" operation typically represent sorted lists from which a histogram could be generated in which position along one (e.g., horizontal) axis indicates abundance number (of subtractant library sequences), and position along another (e.g., vertical) axis indicates identified sequence value (e.g., human or non-human gene type).

The transcript sequences (sorted lists) output from the Tempsubsort and Temptarsort sorting operations are combined during the operation identified as "Cruncher." The "Cruncher" process identifies pairs of corresponding target and subtractant abundance numbers (both representing the same identified sequence value), and divides one by the other to generate a "ratio" value for each pair of corresponding abundance numbers, and then sorts the ratio values in order of decreasing ratio value. The data output from the "Cruncher" operation (the Final Transcript sequence in Fig. 2) is typically a sorted list from which a histogram could be generated in which position along one axis indicates the size of a ratio of abundance numbers (for corresponding identified sequence values from target and subtractant libraries) and position along another axis indicates identified sequence value (e.g., gene type).

Preferably, prior to obtaining a ratio between the two library abundance values, the Cruncher operation also divides each ratio value by the total number of sequences in one or both of the target and subtractant libraries. The resulting lists of "relative" ratio values generated by the Cruncher operation are useful for many medical, scientific, and industrial applications. Also preferably, the output of the Cruncher operation is a set of lists, each list representing a sequence of decreasing ratio values for a different selected subset (e.g. protein family) of database entries.

In one example, the abundance sort program of the invention tabulates for a library the numbers of mRNA transcripts corresponding to each gene identified in a database. These numbers are divided by the total number of clones sampled. The results of the division reflect the relative abundance of the mRNA transcripts in the cell type or tissue from which they were obtained. Obtaining this final data set is referred to herein as "gene transcript image analysis." The resulting subtracted data show exactly what proteins and genes are upregulated and downregulated in highly detailed complexity.

6.6. HUVEC cDNA LIBRARY

Table 2 is an abundance table listing the various gene transcripts in an induced HUVEC library. The transcripts are listed in order of decreasing abundance. This computerized sorting simplifies analysis of the tissue and speeds identification of significant new proteins which are specific to this cell type. This type of endothelial cell lines tissues of the cardiovascular system, and the more that is known about its composition, particularly in response to activation, the more choices of protein targets become available to affect in treating disorders of this tissue, such as the highly prevalent atherosclerosis.

6.7. MONOCYTE-CELL AND MAST-CELL cDNA LIBRARIES

Tables 3 and 4 show truncated comparisons of two libraries. In Tables 3 and 4 the "normal monocytes" are the HMC-1 cells, and the "activated macrophages" are the THP-1 cells pretreated with PMA and activated with LPS. Table 3 lists in descending order of abundance the most abundant gene transcripts for both cell types. With only 15 gene transcripts from each cell type, this table permits quick, qualitative comparison of the most common transcripts. This abundance sort, with its convenient side-by-side display, provides an immediately useful research tool. In this example, this research tool discloses that 1) only one of the top 15 activated macrophage transcripts is found in the top 15 normal monocyte gene transcripts (poly A binding protein); and 2) a new gene transcript (previously unreported in other databases) is relatively highly represented in activated macrophages but is not similarly prominent in normal macrophages. Such a research tool provides researchers with a short-cut to new proteins, such as receptors, cell-surface and intracellular signalling molecules, which can serve as drug targets in commercial drug screening programs. Such a tool could save considerable time over that consumed by a hit and miss discovery program aimed at identifying important proteins in and around cells, because those proteins carrying out everyday cellular functions and

represented as steady state mRNA are quickly eliminated from further characterization.

This illustrates how the gene transcript profiles change with altered cellular function. Those skilled in the art know that the biochemical composition of cells also changes with other functional changes such as cancer, including cancer's various stages, and exposure to toxicity. A gene transcript subtraction profile such as in Table 3 is useful as a first screening tool for such gene expression and protein studies.

6.8. SUBTRACTION ANALYSIS OF NORMAL MONOCYTE-CELL AND ACTIVATED MONOCYTE CELL cDNA LIBRARIES

Once the cDNA data are in the computer, the computer program as disclosed in Table 5 was used to obtain ratios of all the gene transcripts in the two libraries discussed in Example 6.7, and the gene transcripts were sorted by the descending values of their ratios. If a gene transcript is not represented in one library, that gene transcript's abundance is unknown but appears to be less than 1. As an approximation -- and to obtain a ratio, which would not be possible if the unrepresented gene were given an abundance of zero -- genes which are represented in only one of the two libraries are assigned an abundance of 1/2. Using 1/2 for unrepresented clones increases the relative importance of "turned-on" and "turned-off" genes, whose products would be drug candidates. The resulting print-out is called a subtraction table and is an extremely valuable screening method, as is shown by the following data.

Table 4 is a subtraction table, in which the normal monocyte library was electronically "subtracted" from the activated macrophage library. This table highlights most effectively the changes in abundance of the gene transcripts by activation of macrophages. Even among the first 20 gene transcripts listed, there are several unknown gene transcripts. Thus, electronic subtraction is a useful tool with which to assist researchers in identifying much more quickly the basic biochemical changes between two cell types. Such a tool can save universities and pharmaceutical companies which spend billions of dollars on

research valuable time and laboratory resources at the early discovery stage and can speed up the drug development cycle, which in turn permits researchers to set up drug screening programs much earlier. Thus, this research tool
5 provides a way to get new drugs to the public faster and more economically.

Also, such a subtraction table can be obtained for patient diagnosis. An individual patient sample (such as monocytes obtained from a biopsy or blood sample) can be
10 compared with data provided herein to diagnose conditions associated with macrophage activation.

Table 4 uncovered many new gene transcripts (labeled Incyte clones). Note that many genes are turned on in the activated macrophage (i.e., the monocyte had a 0 in the
15 bgfreq column). This screening method is superior to other screening techniques, such as the western blot, which are incapable of uncovering such a multitude of discrete new gene transcripts.

The subtraction-screening technique has also uncovered
20 a high number of cancer gene transcripts (oncogenes rho, ETS2, rab-2 ras, YPT1-related, and acute myeloid leukemia mRNA) in the activated macrophage. These transcripts may be attributed to the use of immortalized cell lines and are inherently interesting for that reason. This screening
25 technique offers a detailed picture of upregulated transcripts including oncogenes, which helps explain why anti-cancer drugs interfere with the patient's immunity mediated by activated macrophages. Armed with knowledge gained from this screening method, those skilled in the art
30 can set up more targeted, more effective drug screening programs to identify drugs which are differentially effective against 1) both relevant cancers and activated macrophage conditions with the same gene transcript profile; 2) cancer alone; and 3) activated macrophage
35 conditions.

Smooth muscle senescent protein (22 kd) was upregulated in the activated macrophage, which indicates that it is a candidate to block in controlling inflammation.

6.9. SUBTRACTION ANALYSIS OF NORMAL LIVER CELLS AND HEPATITIS INFECTED LIVER CELL cDNA LIBRARIES

In this example, rats are exposed to hepatitis virus and maintained in the colony until they show definite signs of hepatitis. Of the rats diagnosed with hepatitis, one half of the rats are treated with a new anti-hepatitis agent (AHA). Liver samples are obtained from all rats before exposure to the hepatitis virus and at the end of AHA treatment or no treatment. In addition, liver samples can be obtained from rats with hepatitis just prior to AHA treatment.

The liver tissue is treated as described in Examples 6.2 and 6.3 to obtain mRNA and subsequently to sequence cDNA. The cDNA from each sample are processed and analyzed for abundance according to the computer program in Table 5. The resulting gene transcript images of the cDNA provide detailed pictures of the baseline (control) for each animal and of the infected and/or treated state of the animals. cDNA data for a group of samples can be combined into a group summary gene transcript profile for all control samples, all samples from infected rats and all samples from AHA-treated rats.

Subtractions are performed between appropriate individual libraries and the grouped libraries. For individual animals, control and post-study samples can be subtracted. Also, if samples are obtained before and after AHA treatment, that data from individual animals and treatment groups can be subtracted. In addition, the data for all control samples can be pooled and averaged. The control average can be subtracted from averages of both post-study AHA and post-study non-AHA cDNA samples. If pre- and post-treatment samples are available, pre- and post-treatment samples can be compared individually (or electronically averaged) and subtracted.

These subtraction tables are used in two general ways. First, the differences are analyzed for gene transcripts which are associated with continuing hepatic deterioration or healing. The subtraction tables are tools to isolate the effects of the drug treatment from the underlying basic pathology of hepatitis. Because hepatitis affects many

parameters, additional liver toxicity has been difficult to detect with only blood tests for the usual enzymes. The gene transcript profile and subtraction provides a much more complex biochemical picture which researchers have
5 needed to analyze such difficult problems.

Second, the subtraction tables provide a tool for identifying clinical markers, individual proteins or other biochemical determinants which are used to predict and/or evaluate a clinical endpoint, such as disease, improvement
10 due to the drug, and even additional pathology due to the drug. The subtraction tables specifically highlight genes which are turned on or off. Thus, the subtraction tables provide a first screen for a set of gene transcript candidates for use as clinical markers. Subsequently,
15 electronic subtractions of additional cell and tissue libraries reveal which of the potential markers are in fact found in different cell and tissue libraries. Candidate gene transcripts found in additional libraries are removed from the set of potential clinical markers. Then, tests of
20 blood or other relevant samples which are known to lack and have the relevant condition are compared to validate the selection of the clinical marker. In this method, the particular physiologic function of the protein transcript need not be determined to qualify the gene transcript as a
25 clinical marker.

6.10. ELECTRONIC NORTHERN BLOT

One limitation of electronic subtraction is that it is difficult to compare more than a pair of images at once. Once particular individual gene products are identified as
30 relevant to further study (via electronic subtraction or other methods), it is useful to study the expression of single genes in a multitude of different tissues. In the lab, the technique of "Northern" blot hybridization is used for this purpose. In this technique, a single cDNA, or a
35 probe corresponding thereto, is labeled and then hybridized against a blot containing RNA samples prepared from a multitude of tissues or cell types. Upon autoradiography,

second set is indicative of one of the biological sequences of the second library. Then the second set of transcript sequences is processed in a programmed computer to generate a second set of identified sequence values, namely the

5 further identified sequence values, each of which is indicative of a sequence annotation and includes a degree of match between one of the biological sequences of the second library and at least one of the reference sequences. The further identified sequence values are processed to

10 generate further final data values indicative of the number of times each further identified sequence value is present in the second library. The final data values from the first specimen and the further identified sequence values from the second specimen are processed to generate ratios

15 of transcript sequences, which indicate the differences in the number of gene transcripts between the two specimens.

In a further embodiment, the method includes quantifying the relative abundance of mRNA in a biological specimen by (a) isolating a population of mRNA transcripts

20 from a biological specimen; (b) identifying genes from which the mRNA was transcribed by a sequence-specific method; (c) determining the numbers of mRNA transcripts corresponding to each of the genes; and (d) using the mRNA transcript numbers to determine the relative abundance of

25 mRNA transcripts within the population of mRNA transcripts.

Also disclosed is a method of producing a gene transcript image analysis by first obtaining a mixture of mRNA, from which cDNA copies are made. The cDNA is inserted into a suitable vector which is used to transfect

30 suitable host strain cells which are plated out and permitted to grow into clones, each clone representing a unique mRNA. A representative population of clones transfected with cDNA is isolated. Each clone in the population is identified by a sequence-specific method

35 which identifies the gene from which the unique mRNA was transcribed. The number of times each gene is identified to a clone is determined to evaluate gene transcript abundance. The genes and their abundances are listed in order of abundance to produce a gene transcript image.

the pattern of expression of that particular gene, one at a time, can be quantitated in all the included samples.

In contrast, a further embodiment of this invention is the computerized form of this process, termed here
5 "electronic northern blot." In this variation, a single gene is queried for expression against a multitude of prepared and sequenced libraries present within the database. In this way, the pattern of expression of any single candidate gene can be examined instantaneously and
10 effortlessly. More candidate genes can thus be scanned, leading to more frequent and fruitfully relevant discoveries. The computer program included as Table 5 includes a program for performing this function, and Table 6 is a partial listing of entries of the database used in
15 the electronic northern blot analysis.

6.11. PHASE I CLINICAL TRIALS

Based on the establishment of safety and effectiveness in the above animal tests, Phase I clinical tests are undertaken. Normal patients are subjected to the usual
20 preliminary clinical laboratory tests. In addition, appropriate specimens are taken and subjected to gene transcript analysis. Additional patient specimens are taken at predetermined intervals during the test. The specimens are subjected to gene transcript analysis as
25 described above. In addition, the gene transcript changes noted in the earlier rat toxicity study are carefully evaluated as clinical markers in the followed patients. Changes in the gene transcript analyses are evaluated as indicators of toxicity by correlation with clinical signs
30 and symptoms and other laboratory results. In addition, subtraction is performed on individual patient specimens and on averaged patient specimens. The subtraction analysis highlights any toxicological changes in the treated patients. This is a highly refined determinant of
35 toxicity. The subtraction method also annotates clinical markers. Further subgroups can be analyzed by subtraction analysis, including, for example, 1) segregation by

occurrence and type of adverse effect; and 2) segregation by dosage.

6.12. GENE TRANSCRIPT IMAGING ANALYSIS IN CLINICAL STUDIES

A gene transcript imaging analysis (or multiple gene
5 transcript imaging analyses) is a useful tool in other
clinical studies. For example, the differences in gene
transcript imaging analyses before and after treatment can
be assessed for patients on placebo and drug treatment.
This method also effectively screens for clinical markers
10 to follow in clinical use of the drug.

6.13. COMPARATIVE GENE TRANSCRIPT ANALYSIS BETWEEN SPECIES

The subtraction method can be used to screen cDNA
libraries from diverse sources. For example, the same cell
types from different species can be compared by gene
15 transcript analysis to screen for specific differences,
such as in detoxification enzyme systems. Such testing
aids in the selection and validation of an animal model for
the commercial purpose of drug screening or toxicological
testing of drugs intended for human or animal use. When
20 the comparison between animals of different species is
shown in columns for each species, we refer to this as an
interspecies comparison, or zoo blot.

Embodiments of this invention may employ databases
such as those written using the FoxBASE programming
25 language commercially available from Microsoft Corporation.
Other embodiments of the invention employ other databases,
such as a random peptide database, a polymer database, a
synthetic oligomer database, or a oligonucleotide database
of the type described in U.S. Patent 5,270,170, issued
30 December 14, 1993 to Cull, et al., PCT International
Application Publication No. WO 9322684, published November
11, 1993, PCT International Application Publication No. WO
9306121, published April 1, 1993, or PCT International
Application Publication No. WO 9119818, published December
35 26, 1991. These four references (whose text is
incorporated herein by reference) include teaching which

may be applied in implementing such other embodiments of the present invention.

All references referred to in the preceding text are hereby expressly incorporated by reference herein.

- 5 Various modifications and variations of the described method and system of the invention will be apparent to those skilled in the art without departing from the scope and spirit of the invention. Although the invention has been described in connection with specific preferred
- 10 embodiments, it should be understood that the invention as claimed should not be unduly limited to such specific embodiments.

TABLE 1

Designations (D)	Distribution (F)	Localization (Z)	Function (R)
E = Exact	C = Non-specific	N = Nuclear	T = Translation
H = Homologous	P = Cell/tissue specific	C = Cytoplasmic	L = Protein processing
O = Other species	U = Unknown	K = Cytoskeleton	R = Ribosomal protein
N = No match		E = Cell surface	O = Oncogene
D = Noncoding gene		Z = Intracellular memb	G = GTP binding ptn
U = Nonreadable		M = Mitochondrial	V = Viral element
R = Repetitive DNA		S = Secreted	Y = Kinase/phosphatase
A = Poly-A only	Species (S)	U = Unknown	A = Tumor antigen related
V = Vector only	H = Human	X = Other	I = Binding proteins
M = Mitochondrial DNA	A = Ape		D = NA-binding /transcription
S = Skip	P = Pig		B = Surface molecule/receptor
I = Match Incyte clone	D = Dog		C = Ca ⁺⁺ binding protein
X = EST match	V = Bovine		S = Ligands/effectors
	B = Rabbit		H = Stress response protein
	R = Rat		E = Enzyme
Library (L)	M = Mouse	Status (I)	F = Ferroprotein
U = U937	S = Hamster	0 = No current interest	P = Protease/inhibitor
M = HMC	C = Chicken	1 = Do primary analysis	Z = Oxidative phosphorylation
T = THP-1	F = Amphibian	2 = Primary analysis done	Q = Sugar metabolism
H = HUVEC	I = Invertebrate	3 = Full length sequence	M = Amino acid metabolism
S = Spleen	Z = Protozoan	4 = Secondary analysis	N = Nucleic acid metabolism
L = Lung	G = Fungi	5 = Tissue northern	W = Lipid metabolism
Y = T & B cell		6 = Obtain full length	K = Structural
A = Adenoid			X = Other
			U = unknown

TABLE 2

Clone numbers 15000 through 20000
 Libraries: HUVEC
 Arranged by ABUNDANCE
 Total clones analyzed: 5000

319 genes, for a total of 1713 Clones

	number	N	c	entry	s	descriptor
1	15365	67		HSRPL41		Riboptn L41
2	15004	65		NCY015004		INCYTE 015004
3	15638	63		NCY015638		INCYTE 015638
4	15390	50		NCY015390		INCYTE 015390
5	15193	47		HSFIB1		Fibronectin
6	15220	47		RRRPL9	R	Riboptn L9
7	15280	47		NCY015280		INCYTE 015280
8	15583	33		M62060		EST HHCH09 (IGR)
9	15662	31		HSACTCGR		Actin, gamma
10	15026	29		NCY015026		INCYTE 015026
11	15279	24		HSEF1AR		Elf 1-alpha
12	15027	23		NCY015027		INCYTE 015027
13	15033	20		NCY015033		INCYTE 015033
14	15198	20		NCY015198		INCYTE 015198
15	15809	20		HSCOLL1		Collagenase
16	15221	19		NCY015221		INCYTE 015221
17	15263	19		NCY015263		INCYTE 015263
18	15290	19		NCY015290		INCYTE 015290
19	15350	18		NCY015350		INCYTE 015350
20	15030	17		NCY015030		INCYTE 015030
21	15234	17		NCY015234		INCYTE 015234
22	15459	16		NCY015459		INCYTE 015459
23	15353	15		NCY015353		INCYTE 015353
24	15378	15		S76965		Ptn kinase inhib
25	15255	14		HUMTHYB4		Thymosin beta-4
26	15401	14		HSLIPCR		Lipocortin I
27	15425	14		HSPOLYAB		Poly-A bp
28	18212	14		HUMTHYMA		Thymosin, alpha
29	18216	14		HSMRP1		Motility relat ptn; MRP-1;CD-9
30	15189	13		HS18D		Interferon induc ptn 1-8D
31	15031	12		HUMFKBP		FK506 bp
32	15306	12		HSH2AZ		Histone H2A
33	15621	12		HUMLEC		Lectin, B-galbp, 14kDa
34	15789	11		NCY015789		INCYTE 015789
35	16578	11		HSRPS11		Riboptn S11
36	16632	11		M61984		EST HHCA13 (IGR)
37	18314	11		NCY018314		INCYTE 018314
38	15367	10		NCY015367		INCYTE 015367
39	15415	10		HSIFNIN1		interferon induc mRNA
40	15633	10		HSLDHAR		Lactate dehydrogenase
41	15813	10		CHKNMHCB		C Myosin heavy chain B
42	18210	10		NCY018210		INCYTE 018210
43	18233	10		HSRPII140		RNA polymerase II
44	18996	10		NCY018996		INCYTE 018996
45	15088	9		HUMFERL		Ferritin, light chain
46	15714	9		NCY015714		INCYTE 015714
47	15720	9		NCY015720		INCYTE 015720
48	15863	9		NCY015863		INCYTE 015863
49	16121	9		HSET		Endothelin
50	18252	9		NCY018252		INCYTE 018252
51	15351	8		HUMALBP		Lipid bp, adipocyte
52	15370	8		NCY015370		INCYTE 015370

TABLE 2 Con't

	number	N	c	entry	s	descriptor
53	15670	8		BTCIASHI	V	NADH-ubiq oxidoreductase
54	15795	8		NCY015795		INCYTE 015795
55	16245	8		NCY016245		INCYTE 016245
56	18262	8		NCY018262		INCYTE 018262
57	18321	8		HSRPL17		Riboptn L17
58	15126	7		XLRPL1BRF		Riboptn L1
59	15133	7		HSAC07		Actin, beta
60	15245	7		NCY015245		INCYTE 015245
61	15288	7		NCY015288		INCYTE 015288
62	15294	7		HSGAPDR		G-3-PD
63	15442	7		HUMLAMB		Laminin receptor, 54kDa
64	15485	7		HSNGMRNA		Uracil DNA glycosylase
65	16646	7		NCY016646		INCYTE 016646
66	18003	7		HUMPAIA		Plsmnogen activ gene
67	15032	6		HUMUB		Ubiquitin
68	15267	6		HSRPS8		Riboptn S8
69	15295	6		NCY015295		INCYTE 015295
70	15458	6		RNRPS10R	R	Riboptn S10
71	15832	6		RSGALEM	R	UDP-galactose epimerase
72	15928	6		HUMAPOJ		Apolipoptn J
73	16598	6		HUMTBMM40		Tubulin, beta
74	18218	6		NCY018218		INCYTE 018218
75	18499	6		HSP27		Hydrophobic ptn p27
76	18963	6		NCY018963		INCYTE 018963
77	18997	6		NCY018997		INCYTE 018997
78	15432	5		HSAGALAR		Galactosidase A, alpha
79	15475	5		NCY015475		INCYTE 015475
80	15721	5		NCY015721		INCYTE 015721
81	15865	5		NCY015865		INCYTE 015865
82	16270	5		NCY016270		INCYTE 016270
83	16886	5		NCY016886		INCYTE 016886
84	18500	5		NCY018500		INCYTE 018500
85	18503	5		NCY018503		INCYTE 018503
86	19672	5		RRRPL34	R	Riboptn L34
87	15086	4		XLRPL1AR	F	Riboptn L1a
88	15113	4		HUMIFNWRS		tRNA synthetase, trp
89	15242	4		NCY015242		INCYTE 015242
90	15249	4		NCY015249		INCYTE 015249
91	15377	4		NCY015377		INCYTE 015377
92	15407	4		NCY015407		INCYTE 015407
93	15473	4		NCY015473		INCYTE 015473
94	15588	4		HSRPS12		Riboptn S12
95	15684	4		HSEF1G		Elf 1-gamma
96	15782	4		NCY015782		INCYTE 015782
97	15916	4		HSRPS18		Riboptn S18
98	15930	4		NCY015930		INCYTE 015930
99	16108	4		NCY016108		INCYTE 016108
100	16133	4		NCY016133		INCYTE 016133

NORMAL MONONCYTE VS. ACTIVATED MACROPHAGE

Top 15 Most Abundant Genes

NORMAL

- 1 Elongation factor-1 alpha
- 2 Ribosomal phosphoprotein
- 3 Ribosomal protein S8 homolog
- 4 Beta-Globin
- 5 Ferritin H chain
- 6 Ribosomal protein L7
- 7 Nucleoplasmin
- 8 Ribosomal protein S20 homolog
- 9 Transferrin receptor
- 10 Poly-A binding protein
- 11 Translationally controlled tumor ptn
- 12 Ribosomal protein S25
- 13 Signal recognition particle SRP9
- 14 Histone H2A.Z
- 15 Ribosomal protein Ke-3

ACTIVATED

- Interleukin-1 beta
- Macrophage inflammatory protein-1
- Interleukin-8
- Lymphocyte activation gene
- Elongation factor-1 alpha
- Beta actin
- Rantes T-cell specific protein
- Poly A binding protein
- Osteopontin; nephropontin
- Tumor Necrosis Factor-alpha
- INCYTE clone 011050
- Cu/Zn superoxide dismutase
- Adenylate cyclase (yeast homolog)
- NGF-related B cell activation molecule
- Protease Nexin-1, glial-derived

TABLE 3

TABLE 4

Libraries: THP-1
 Subtracting: HMC
 Sorted by ABUNDANCE
 Total clones analyzed: 7375

1057 genes, for a total of 2151 clones

number	entry	s descriptor	bgfreq	rfend	ratio
10022	HUMIL1	IL 1-beta	0	131	262.00
10036	HSMDNCF	IL-8	0	119	238.00
10089	HSLAG1CDN	Lymphocyte activ gene	0	71	142.00
10060	HUMTCSM	RANTES	0	23	46.000
10003	HUMMIP1A	MIP-1	3	121	40.333
10689	HSOP	Osteopontin	0	20	40.000
11050	NCY011050	INCYTE 011050	0	17	34.000
10937	HSTNFR	TNF-alpha	0	17	34.000
10176	HSSOD	Superoxide dismutase	0	14	28.000
10886	HSCDW40	B-cell activ,NGF-relat	0	10	20.000
10186	HUMAPR	Early resp PMA-induc	0	9	18.000
10967	HUMGDN	PN-1, glial-deriv	0	9	18.000
11353	NCY011353	INCYTE 011353	0	8	16.000
10298	NCY010298	INCYTE 010298	0	7	14.000
10215	HUM4COLA	Collagenase, type IV	0	6	12.000
10276	NCY010276	INCYTE 010276	0	6	12.000
10488	NCY010488	INCYTE 010488	0	6	12.000
11138	NCY011138	INCYTE 011138	0	6	12.000
10037	HUMCAPPRO	Adenylate cyclase	1	10	10.000
10840	HUMADCY	Adenylate cyclase	0	5	10.000
10672	HSCD44E	Cell adhesion glptn	0	5	10.000
12837	HUMCYCLOX	Cyclooxygenase-2	0	5	10.000
10001	NCY010001	INCYTE 010001	0	5	10.000
10005	NCY010005	INCYTE 010005	0	5	10.000
10294	NCY010294	INCYTE 010294	0	5	10.000
10297	NCY010297	INCYTE 010297	0	5	10.000
10403	NCY010403	INCYTE 010403	0	5	10.000
10699	NCY010699	INCYTE 010699	0	5	10.000
10966	NCY010966	INCYTE 010966	0	5	10.000
12092	NCY012092	INCYTE 012092	0	5	10.000
12549	HSRHOB	Oncogene rho	0	5	10.000
10691	HUMARF1BA	ADP-ribosylation fctr	0	4	8.000
12106	HSADSS	Adenylosuccinate synthetase	0	4	8.000
10194	HSCATHL	Cathepsin L	0	4	8.000
10479	CLMCYCA	I Cyclin A	0	4	8.000
10031	NCY010031	INCYTE 010031	0	4	8.000
10203	NCY010203	INCYTE 010203	0	4	8.000
10288	NCY010288	INCYTE 010288	0	4	8.000
10372	NCY010372	INCYTE 010372	0	4	8.000
10471	NCY010471	INCYTE 010471	0	4	8.000
10484	NCY010484	INCYTE 010484	0	4	8.000
10859	NCY010859	INCYTE 010859	0	4	8.000
10890	NCY010890	INCYTE 010890	0	4	8.000
11511	NCY011511	INCYTE 011511	0	4	8.000
11868	NCY011868	INCYTE 011868	0	4	8.000
12820	NCY012820	INCYTE 012820	0	4	8.000
10133	HSI1RAP	IL-1 antagonist	0	4	8.000
10516	HUMP2A	Phosphatase, regul 2A	0	4	8.000
11063	HUMB94	TNF-induc response	0	4	8.000
11140	HSHB15RNA	HB15 gene; new Ig	0	3	6.000
10788	NCY001713	INCYTE 001713	0	3	6.000
10033	NCY010033	INCYTE 010033	0	3	6.000
10035	NCY010035	INCYTE 010035	0	3	6.000
10084	NCY010084	INCYTE 010084	0	3	6.000
10236	NCY010236	INCYTE 010236	0	3	6.000
10383	NCY010383	INCYTE 010383	0	3	6.000

TABLE 4 Con't

number	entry	s descriptor	bgfreq	rfend	ratio
10450	NCY010450	INCYTE 010450	0	3	6.000
10470	NCY010470	INCYTE 010470	0	3	6.000
10504	NCY010504	INCYTE 010504	0	3	6.000
10507	NCY010507	INCYTE 010507	0	3	6.000
10598	NCY010598	INCYTE 010598	0	3	6.000
10779	NCY010779	INCYTE 010779	0	3	6.000
10909	NCY010909	INCYTE 010909	0	3	6.000
10976	NCY010976	INCYTE 010976	0	3	6.000
10985	NCY010985	INCYTE 010985	0	3	6.000
11052	NCY011052	INCYTE 011052	0	3	6.000
11068	NCY011068	INCYTE 011068	0	3	6.000
11134	NCY011134	INCYTE 011134	0	3	6.000
11136	NCY011136	INCYTE 011136	0	3	6.000
11191	NCY011191	INCYTE 011191	0	3	6.000
11219	NCY011219	INCYTE 011219	0	3	6.000
11386	NCY011386	INCYTE 011386	0	3	6.000
11403	NCY011403	INCYTE 011403	0	3	6.000
11460	NCY011460	INCYTE 011460	0	3	6.000
11618	NCY011618	INCYTE 011618	0	3	6.000
11686	NCY011686	INCYTE 011686	0	3	6.000
12021	NCY012021	INCYTE 012021	0	3	6.000
12025	NCY012025	INCYTE 012025	0	3	6.000
12320	NCY012320	INCYTE 012320	0	3	6.000
12330	NCY012330	INCYTE 012330	0	3	6.000
12853	NCY012853	INCYTE 012853	0	3	6.000
14386	NCY014386	INCYTE 014386	0	3	6.000
14391	NCY014391	INCYTE 014391	0	3	6.000

TABLE 5

```

* Master menu for SUBTRACTION output
SET TALK OFF
SET SAFETY OFF
SET EXACT ON
SET TYPEAHEAD TO 0
CLEAR
SET DEVICE TO SCREEN
USE "SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf"
GO TOP
STORE NUMBER TO INITIATE
GO BOTTOM
STORE NUMBER TO TERMINATE
STORE : TO Target1
STORE : TO Target2
STORE : TO Target3
STORE : TO Object1
STORE : TO Object2
STORE : TO Object3
STORE 0 TO ANAL
STORE 0 TO EMATCH
STORE 0 TO HMATCH
STORE 0 TO OMATCH
STORE 0 TO IMATCH
STORE 0 TO PTF
STORE 1 TO BAIL
DO WHILE .T.
* Program.: Subtraction 2.fmt
* Date..... 10/11/94
* Version.: FoxBASE+/Mac, revision 1.10
* Notes..... Format file Subtraction 2
*
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",9 COLOR 0,0,0,
@ PIXELS 75,120 TO 178,241 STYLE 3871 COLOR 0,0,-1,24610,-1,8947
@ PIXELS 27,134 SAY "Subtraction Menu" STYLE 65536 FONT "Geneva",274 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 117,126 GET EMATCH STYLE 65536 FONT "Chicago",12 PICTURE "e*c Exact " SIZE 15,62 CO
@ PIXELS 135,126 GET HMATCH STYLE 65536 FONT "Chicago",12 PICTURE "e*c Homologous" SIZE 15,1
@ PIXELS 153,126 GET OMATCH STYLE 65536 FONT "Chicago",12 PICTURE "e*c Other spc" SIZE 15,84
@ PIXELS 90,152 SAY "Matches:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 171,126 GET Imatch STYLE 65536 FONT "Chicago",12 PICTURE "e*c Incyte" SIZE 15,65 CO
@ PIXELS 252,137 GET initiate STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 252,236 GET terminate STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 252,35 SAY "Include clones:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 252,215 SAY "->" STYLE 65536 FONT "Geneva",14 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 198,126 GET PTF STYLE 65536 FONT "Chicago",12 PICTURE "e*c Print to file" SIZE 15,9
@ PIXELS 90,9 TO 181,109 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 90,288 TO 181,397 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 81,296 SAY "Background:" STYLE 65536 FONT "Geneva",270 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 45,135 GET ANAL STYLE 65536 FONT "Chicago",12 PICTURE "e*R Overall;Function" SIZE 4
@ PIXELS 81,26 SAY "Target:" STYLE 65536 FONT "Geneva",270 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 108,20 GET target1 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 135,20 GET target2 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 162,20 GET target3 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 108,299 GET object1 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 135,299 GET object2 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 162,299 GET object3 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 276,324 GET Bail STYLE 65536 FONT "Chicago",12 PICTURE "e*R Run;Bail out" SIZE 4112
*
* EOF: Subtraction.2.fmt
READ
IF Bail=2
CLEAR
CLOSE DATABASES
USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
SET SAFETY ON
SCREEN 1 OFF
RETURN

```



```

ENDIF
STORE VAL(SYS(2)) TO STARTIME
STORE UPPER(Target1) TO Target1
STORE UPPER(Target2) TO Target2
STORE UPPER(Target3) TO Target3
STORE UPPER(Object1) TO Object1
STORE UPPER(Object2) TO Object2
STORE UPPER(Object3) TO Object3
clear
SET TALK ON
GAP = TERMINATE-INITIATE+1
GO INITIATE
COPY NEXT GAP FIELDS NUMBER, library, D, F, Z, R, ENTRY, S, DESCRIPTOR, START, RFEND, I TO TEMPNUM
USE TEMPNUM
COUNT TO TOT
COPY TO TEMPRED FOR D='E'.OR.D='O'.OR.D='H'.OR.D='N'.OR.D='I'
USE TEMPRED

IF Ematch=0 .AND. Hmatch=0 .AND. Omatch=0 .AND. IMATCH=0
COPY TO TEMPDESIG
ELSE
COPY STRUCTURE TO TEMPDESIG
USE TEMPDESIG
  IF Ematch=1
    APPEND FROM TEMPNUM FOR D='E'
  ENDIF
  IF Hmatch=1
    APPEND FROM TEMPNUM FOR D='H'
  ENDIF
  IF Omatch=1
    APPEND FROM TEMPNUM FOR D='O'
  ENDIF
  IF Imatch=1
    APPEND FROM TEMPNUM FOR D='I'.OR.D='X'
    *.OR.D='N'
  ENDIF
ENDIF
COUNT TO STARTOT

COPY STRUCTURE TO TEMPLIB
USE TEMPLIB
  APPEND FROM TEMPDESIG FOR library=UPPER(target1)
  IF target2<>'
    APPEND FROM TEMPDESIG FOR library=UPPER(target2)
  ENDIF
  IF target3<>'
    APPEND FROM TEMPDESIG FOR library=UPPER(target3)
  ENDIF
COUNT TO ANALTOT

USE TEMPDESIG
COPY STRUCTURE TO TEMPSUB
USE TEMPSUB
  APPEND FROM TEMPDESIG FOR library=UPPER(Object1)
  IF target2<>'
    APPEND FROM TEMPDESIG FOR library=UPPER(Object2)
  ENDIF
  IF target3<>'
    APPEND FROM TEMPDESIG FOR library=UPPER(Object3)
  ENDIF
COUNT TO SUBTRACTOT
SET TALK OFF
*****
* COMPRESSION SUBROUTINE A
? 'COMPRESSING QUERY LIBRARY'
USE TEMPLIB

```

```

SORT ON ENTRY, NUMBER TO LIBSORT
USE LIBSORT
COUNT TO IDGENE
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= IDGENE
    PACK
    COUNT TO AUNIQUE
    SW2=1
  LOOP
ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
STORE D TO DESIGA
SW = 0
DO WHILE SW=0 TEST
  SKIP
  STORE ENTRY TO TESTB
  STORE D TO DESIGB
  IF TESTA = TESTB.AND.DESIGA=DESIGB
    DELETE
    DUP = DUP+1
  LOOP
ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
SORT ON RFEND/D, NUMBER TO TEMPTARSORT
USE TEMPTARSORT
*REPLACE ALL START WITH RFEND/IDGENE*10000
COUNT TO TEMPTARCO
*****
* COMPRESSION SUBROUTINE B
? 'COMPRESSING TARGET LIBRARY'
USE TEMPSUB
SORT ON ENTRY, NUMBER TO SUBSORT
USE SUBSORT
COUNT TO SUBGENE
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= SUBGENE
    PACK
    COUNT TO BUNIQUE
    SW2=1
  LOOP
ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
STORE D TO DESIGA
SW = 0
DO WHILE SW=0 TEST
  SKIP
  STORE ENTRY TO TESTB
  STORE D TO DESIGB
  IF TESTA = TESTB.AND.DESIGA=DESIGB

```

```

DELETE
DUP = DUP+1
LOOP
ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP :
ENDDO ROLL
SORT ON RFEND/D,NUMBER TO TEMPSUBSORT
USE TEMPSUBSORT
*REPLACE ALL START WITH RFEND/IDGENE*10000
COUNT TO TEMPSUBCO
*****
*FUSION ROUTINE
? 'SUBTRACTING LIBRARIES'
USE SUBTRACTION
COPY STRUCTURE TO CRUNCHER
SELECT 2
USE TEMPSUBSORT
SELECT 1
USE CRUNCHER
APPEND FROM TEMPTARSORT
COUNT TO BAILOUT
MARK = 0

DO WHILE .T.
SELECT 1
MARK = MARK+1
IF MARK>BAILOUT
EXIT
ENDIF
GO MARK
STORE ENTRY TO SCANNER
SELECT 2
LOCATE FOR ENTRY=SCANNER
IF FOUND()
STORE RFEND TO BIT1
STORE RFEND TO BIT2
ELSE
STORE 1/2 TO BIT1
STORE 0 TO BIT2
ENDIF
SELECT 1
REPLACE BGFREQ WITH BIT2
REPLACE ACTUAL WITH BIT1
LOOP
ENDDO

SELECT 1
REPLACE ALL RATIO WITH RFEND/ACTUAL
? 'DOING FINAL SORT BY RATIO'
SORT ON RATIO/D,BGFREQ/D,DESCRIPTOR TO FINAL
USE FINAL
*****
set talk off
DO CASE
CASE PTF=0
SET DEVICE TO PRINT
SET PRINT ON
EJECT
CASE PTF=1
SET ALTERNATE TO "Adenoid.Patent.Figures.Subtraction.txt"

```

```
SET ALTERNATE ON
ENDCASE
```

```
STORE VAL(SYS(2)) TO FINTIME
IF FINTIME<STARTIME
STORE FINTIME+86400 TO FINTIME
ENDIF
STORE FINTIME - STARTIME TO COMPSEC
STORE COMPSEC/60 TO COMPMIN
```

```
*****
```

```
SET MARGIN TO 10
```

```
@1,1 SAY "Library Subtraction Analysis" STYLE 65536 FONT "Geneva",274 COLOR 0,0,0,-1,-1,-1
```

```
?
```

```
?
```

```
?
```

```
?
```

```
? date()
```

```
?? ' ' .
```

```
?? TIME()
```

```
? 'Clone numbers '
```

```
?? STR(INITIATE,5,0)
```

```
?? ' through ' .
```

```
?? STR(TERMINATE,6,0)
```

```
? 'Libraries: '
```

```
? Target1
```

```
IF Target2<>'
```

```
?? ' , ' .
```

```
?? Target2
```

```
ENDIF
```

```
IF Target3<>'
```

```
?? ' , ' .
```

```
?? Target3
```

```
ENDIF
```

```
? 'Subtracting:
```

```
? Object1
```

```
IF Object2<>'
```

```
?? ' , ' .
```

```
?? Object2
```

```
ENDIF
```

```
IF Object3<>'
```

```
?? ' , ' .
```

```
?? Object3
```

```
ENDIF
```

```
? 'Designations: '
```

```
IF Ematch=0 .AND. Hmatch=0 .AND. Omatch=0 .AND. IMATCH=0
```

```
?? 'All'
```

```
ENDIF
```

```
IF Ematch=1
```

```
?? 'Exact, '
```

```
ENDIF
```

```
IF Hmatch=1
```

```
?? 'Human, '
```

```
ENDIF
```

```
IF Omatch=1
```

```
?? 'Other sp. '
```

```
ENDIF
```

```
IF Imatch=1
```

```
?? 'INCYTE'
```

```
ENDIF
```

```
IF ANAL=1
```

```
? 'Sorted by ABUNDANCE'
```

```
ENDIF
```

```
IF ANAL=2
```

```
? 'Arranged by FUNCTION'
```

```
ENDIF
```

```

? 'Total clones represented: '
?? STR(TOT,5,0)
? 'Total clones analyzed: '
?? STR(STARTOT,5,0)
? 'Total computation time: '
?? STR(COMPMIN,5,2)
?? ' minutes'
?
? 'd = designation   f = distribution   z = location   r = function   s = species   i = inte
?
*****
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',9 COLOR 0,0,0,
DO CASE
CASE ANAL=1
?? STR(AUNIQUE,4,0)
?? ' genes, for a total of '
?? STR(ANALTOT,4,0)
?? ' clones'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I
SET PRINT OFF
CLOSE DATABASES
USE 'SmartGuy:FoxBASE+/Mac:fox files:clones.dbf'

CASE ANAL=2
* arrange/function
SET PRINT ON
SET HEADING ON
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',268 COLOR 0
?
? '
          BINDING PROTEINS'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Surface molecules and receptors:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='B'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Calcium-binding proteins:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='C'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Ligands and effectors:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='S'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Other binding proteins:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='I'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',268 COLOR 0
? '
          ONCOGENES'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'General oncogenes:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='O'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'GTP-binding proteins:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='G'

```

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? "Viral elements:"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='V'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? "Kinases and Phosphatases:"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='Y'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? "Tumor-related antigens:"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='A'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",268 COLOR 0
 ? "PROTEIN SYNTHETIC MACHINERY PROTEINS:"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? "Transcription and Nucleic Acid-binding proteins:"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='D'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? "Translation:"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='T'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? "Ribosomal proteins:"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='R'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? "Protein processing:"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='L'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",268 COLOR 0
 ? "ENZYMES:"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? "Ferroproteins:"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='F'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? "Proteases and inhibitors:"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='P'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? "Oxidative phosphorylation:"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='Z'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? "Sugar metabolism:"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='Q'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? "Amino acid metabolism:"

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,

list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='M'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? 'Nucleic acid metabolism:'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='N'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? 'Lipid metabolism:'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='W'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? 'Other enzymes:'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='E'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",268 COLOR 0
 ?
 ? ' MISCELLANEOUS CATEGORIES'
 ?

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? 'Stress response:'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='H'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? 'Structural:'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='K'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? 'Other clones:'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='X'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Helvetica",265 COLOR 0
 ? 'Clones of unknown function:'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
 list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='U'

ENDCASE

DO "Test print.prg"
 SET PRINT OFF
 SET DEVICE TO SCREEN
 CLOSE DATABASES
 ERASE TEMPLIB.DBF
 ERASE TEMPNUM.DBF
 ERASE TEMPDESIG.DBF
 SET MARGIN TO 0
 CLEAR
 LOOP
 ENDDO

```

*Northern (single), version 11-25-94
close databases
SET TALK OFF
SET PRINT OFF
SET EXACT OFF
CLEAR
STORE ' ' TO Eobject
STORE ' ' TO Dobject
STORE 0 TO Numb
STORE 0 TO Zog
STORE 1 TO Bail
DO WHILE .T.
* Program.: Northern (single).fmt
* Date.....: 8/ 8/94
* Version...: FoxBASE+/Mac, revision 1.10
* Notes.....: Format file Northern (single)
*
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",12 COLOR 0,0,0
@ PIXELS 15,81 TO 46,397 STYLE 28447 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 89,79 TO 192,422 STYLE 28447 COLOR 0,0,0,-25600,-1,-1
@ PIXELS 115,98 SAY "Entry #:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 115,173 GET Eobject STYLE 0 FONT "Geneva",12 SIZE 15,142 COLOR 0,0,0,-1,-1,-1
@ PIXELS 145,89 SAY "Description" STYLE 65536 FONT "Geneva",12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 145,173 GET Dobject STYLE 0 FONT "Geneva",12 SIZE 15,241 COLOR 0,0,0,-1,-1,-1
@ PIXELS 35,89 SAY "Single Northern search screen" STYLE 65536 FONT "Geneva",274 COLOR 0,0,-
@ PIXELS 220,162 GET Bail STYLE 65536 FONT "Chicago",12 PICTURE "Q*R Continue;Bail out" SIZE
@ PIXELS 175,98 SAY "Clone #:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 175,173 GET Numb STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,0,-1,-1,-1
@ PIXELS 80,152 SAY "Enter any ONE of the following:" STYLE 65536 FONT "Geneva",12 COLOR -1,
*
* EOF: Northern (single).fmt
READ
IF Bail=2
CLEAR
screen 1 off
RETURN
ENDIF
USE "SmartGuy\FoxBASE+/Mac\Fox files\Lookup.dbf"
SET TALK ON

IF Eobject<>'
STORE UPPER(Eobject) to Eobject
SET SAFETY OFF
SORT ON Entry TO "Lookup entry.dbf"
SET SAFETY ON
USE "Lookup entry.dbf"
LOCATE FOR Look=Eobject
IF .NOT.FOUND()
CLEAR
LOOP
ENDIF
BROWSE
STORE Entry TO Searchval
CLOSE DATABASES
ERASE "Lookup entry.dbf"
ENDIF

IF Dobject<>'
SET EXACT OFF
SET SAFETY OFF
SORT ON descriptor TO "Lookup descriptor.dbf"
SET SAFETY On
USE "Lookup descriptor.dbf"
LOCATE FOR UPPER(TRIM(descriptor))=UPPER(TRIM(Dobject))
IF .NOT.FOUND()
CLEAR

```

```

LOOP
ENDIF
BROWSE
STORE Entry TO Searchval
CLOSE DATABASES
ERASE 'Lookup descriptor.dbf'
SET EXACT ON
ENDIF

IF Numb<0
USE 'SmartGuy\FoxBASE+Mac\Fox files\clones.dbf'
GO Numb
BROWSE
STORE Entry TO Searchval
ENDIF

CLEAR
? 'Northern analysis for entry '
?? Searchval
?
? 'Enter Y to proceed'
WAIT TO OK
CLEAR
IF UPPER(OK)='Y'
screen 1 off
RETURN
ENDIF

* COMPRESSION SUBROUTINE FOR Library.dbf
? 'Compressing the Libraries file now...'
USE 'SmartGuy\FoxBASE+Mac\Fox files\libraries.dbf'
SET SAFETY OFF
SORT ON library TO 'Compressed libraries.dbf'
* FOR entered=0
SET SAFETY ON
USE 'Compressed libraries.dbf'
DELETE FOR entered=0
PACK
COUNT TO TOT
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
    PACK
    SW2=1
    LOOP
  ENDIF
  GO MARK1
  STORE library TO TESTA
  SKIP
  STORE Library TO TESTB
  IF TESTA = TESTB
    DELETE
  ENDIF
  MARK1 = MARK1+1
  LOOP
ENDDO ROLL

* Northern analysis
CLEAR
? 'Doing the northern now...'
SET TALK ON
USE 'SmartGuy\FoxBASE+Mac\Fox files\clones.dbf'
SET SAFETY OFF
COPY TO 'Hits.dbf' FOR entry=searchval
SET SAFETY ON

```

```

* MASTER ANALYSIS 3, VERSION 12-9-94
* Master menu for analysis output
CLOSE DATABASES
SET TALK OFF
SET SAFETY OFF
CLEAR
SET DEVICE TO SCREEN
SET DEFAULT TO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:"
USE "SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf"
GO TOP
STORE NUMBER TO INITIATE
GO BOTTOM
STORE NUMBER TO TERMINATE
STORE 0 TO ENTIRE
STORE 0 TO CONDEN
STORE 0 TO ANAL
STORE 0 TO EMATCH
STORE 0 TO HMATCH
STORE 0 TO OMATCH
STORE 0 TO IMATCH
STORE 0 TO XMATCH
STORE 0 TO PRINTON
STORE 0 TO PTF
DO WHILE .T.
* Program.: Master analysis.fmt
* Date.....: 12/ 9/94
* Version.: FoxBASE+/Mac, revision 1.10
* Notes.....: Format file Master analysis
*
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",9 COLOR 0,0,0,
@ PIXELS 39,255 TO 277,430 STYLE 28447 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 75,120 TO 178,241 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 27,98 SAY "Customized Output Menu" STYLE 65536 FONT "Geneva",274 COLOR 0,0,-1,-1,-1
@ PIXELS 45,54 GET conden STYLE 65536 FONT "Chicago",12 PICTURE "@*C Condensed format" SIZE
@ PIXELS 54,261 GET anal STYLE 65536 FONT "Chicago",12 PICTURE "@*RV Sort/number;Sort/entry;"
@ PIXELS 117,126 GET EMATCH STYLE 65536 FONT "Chicago",12 PICTURE "@*C Exact " SIZE 15,62 CO
@ PIXELS 135,126 GET HMATCH STYLE 65536 FONT "Chicago",12 PICTURE "@*C Homologous" SIZE 15,1
@ PIXELS 153,126 GET OMATCH STYLE 65536 FONT "Chicago",12 PICTURE "@*C Other spc" SIZE 15,84
@ PIXELS 90,152 SAY "Matches:" STYLE 65536 FONT "Geneva",268 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 63,54 GET PRINTON STYLE 65536 FONT "Chicago",12 PICTURE "@*C Include clone listing"
@ PIXELS 171,126 GET Imatch STYLE 65536 FONT "Chicago",12 PICTURE "@*C Incyte" SIZE 15,65 CO
@ PIXELS 252,146 GET initiate STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 270,146 GET terminate STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 234,134 SAY "Include clones " STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 270,125 SAY "->" STYLE 65536 FONT "Geneva",14 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 198,126 GET PTF STYLE 65536 FONT "Chicago",12 PICTURE "@*C Print to file" SIZE 15,9
@ PIXELS 189,0 TO 257,120 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 209,8 SAY "Library selection" STYLE 65536 FONT "Geneva",266 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 227,18 GET ENTIRE STYLE 65536 FONT "Chicago",12 PICTURE "@*RV All;Selected" SIZE 16
*
* EOF: Master analysis.fmt
READ
IF ANAL=9
CLEAR
CLOSE DATABASES
ERASE TEMPMASTER.DBF
USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
SET SAFETY ON
SCREEN 1 OFF
RETURN
ENDIF
clear
? INITIATE
? TERMINATE
? CONDEN
? ANAL

```

```

? ematch
? Hmatch
? Omatch
? IMATCH
SET TALK ON
IF ENTIRE=2
USE "Unique libraries.dbf"
REPLACE ALL i WITH ' '
BROWSE FIELDS i, libname, library, total, entered AT 0,0
ENDIF
USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
*COPY TO TEMPNUM FOR NUMBER>=INITIATE.AND.NUMBER<=TERMINATE
*USE TEMPNUM
COPY STRUCTURE TO TEMPLIB
USE TEMPLIB
IF ENTIRE=1
APPEND FROM "SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf"
ENDIF
IF ENTIRE=2
USE "Unique libraries.dbf"
COPY TO SELECTED FOR UPPER(i)='Y'
USE SELECTED
STORE RECCOUNT() TO STOPIT
MARK=1
DO WHILE .T.
IF MARK>STOPIT
CLEAR
EXIT
ENDIF
USE SELECTED
GO MARK
STORE library TO THISONE
? 'COPYING '
?? THISONE
USE TEMPLIB
APPEND FROM "SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf" FOR library=THISONE
STORE MARK+1 TO MARK
LOOP
ENDDO
ENDIF
USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
COUNT TO STARTOT
COPY STRUCTURE TO TEMPDESIG
USE TEMPDESIG
IF Ematch=0 .AND. Hmatch=0 .AND. Omatch=0 .AND. IMATCH=0
APPEND FROM TEMPLIB
ENDIF
IF Ematch=1
APPEND FROM TEMPLIB FOR D='E'
ENDIF
IF Hmatch=1
APPEND FROM TEMPLIB FOR D='H'
ENDIF
IF Omatch=1
APPEND FROM TEMPLIB FOR D='O'
ENDIF
IF Imatch=1
APPEND FROM TEMPLIB FOR D='I'.OR.D='X'.OR.D='N'
ENDIF
IF Xmatch=1
APPEND FROM TEMPLIB FOR D='X'
ENDIF
COUNT TO ANALTOT
set talk off
*****
DO CASE

```

```

CASE PTF=0
SET DEVICE TO PRINT
SET PRINT ON
EJECT
CASE PTF=1
SET ALTERNATE TO "Total function sort.txt"
*SET ALTERNATE TO "H and O function sort.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Abundance sort.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Abundance con.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Function sort.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Distribution sort.txt"
*SET ALTERNATE TO "Shear stress HUVEC 1:Clone list.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Location sort.txt"
SET ALTERNATE ON
ENDCASE
*****
IF PRINTE=1
@1,30 SAY "Database Subset Analysis" STYLE 65536 FONT "Geneva",274 COLOR 0,0,0,-1,-1,-1
ENDIF
?
?
?
?
? date()
?? '
?? TIME()
? 'Clone numbers '
?? STR(INITIATE,6,0)
?? ' through '
?? STR(TERMINATE,6,0)
? 'Libraries: '
IF ENTIRE=1
? 'All libraries'
ENDIF
IF ENTIRE=2
MARK=1
DO WHILE .T.
IF MARK>STOPIT
EXIT
ENDIF
USE SELECTED
GO MARK
? '
?? TRIM(libname)
STORE MARK+1 TO MARK
LOOP
ENDDO
ENDIF
? 'Designations: '
IF Ematch=0 .AND. Hmatch=0 .AND. Omatch=0 .AND. IMATCH=0
?? 'All'
ENDIF
IF Ematch=1
?? 'Exact,'
ENDIF
IF Hmatch=1
?? 'Human,'
ENDIF
IF Omatch=1
?? 'Other.sp.'
ENDIF
IF Imatch=1
?? 'INCYTE'
ENDIF
IF Xmatch=1
?? 'EST'

```



```

ENDIF
IF CONDEN=1
? 'Condensed format analysis'
ENDIF
IF ANAL=1
? 'Sorted by NUMBER'
ENDIF
IF ANAL=2
? 'Sorted by ENTRY'
ENDIF
IF ANAL=3
? 'Arranged by ABUNDANCE'
ENDIF
IF ANAL=4
? 'Sorted by INTEREST'
ENDIF
IF ANAL=5
? 'Arranged by LOCATION'
ENDIF
IF ANAL=6
? 'Arranged by DISTRIBUTION'
ENDIF
IF ANAL=7
? 'Arranged by FUNCTION'
ENDIF
? 'Total clones represented: '
?? STR(STARTOT,6,0)
? 'Total clones analyzed: '
?? STR(ANAL/TOT,6,0)
?
? 'l = library    d = designation    f = distribution    z = location    r = function    c = cer
?
*****
USE TEMPDESIG
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
DO CASE
CASE ANAL=1
* sort/number
SET HEADING ON
IF CONDEN=1
SORT TO TEMP1 ON ENTRY,NUMBER
DO "COMPRESSION number.PRG"
ELSE
SORT TO TEMP1 ON NUMBER
USE TEMP1
list off fields number,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR
*list off fields number,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,REND,INIT,I
CLOSE DATABASES
ERASE TEMP1.DBF
ENDIF

CASE ANAL=2
* sort/DESCRIPTOR
SET HEADING ON
*SORT TO TEMP1 ON DESCRIPTOR,ENTRY,NUMBER/S for D='E'.OR.D='H'.OR.D='O'.OR.D='X'.OR.D='I'
*SORT TO TEMP1 ON ENTRY,DESCRIPTOR,NUMBER/S for D='E'.OR.D='H'.OR.D='O'.OR.D='X'.OR.D='I'
SORT TO TEMP1 ON ENTRY,START/S for D='E'.OR.D='H'.OR.D='O'.OR.D='X'.OR.D='I'
IF CONDEN=1
DO "COMPRESSION entry.PRG"
ELSE
USE TEMP1
list off fields number,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,REND,INIT,I
CLOSE DATABASES
ERASE TEMP1.DBF
ENDIF

```

```

CASE ANAL=3
* sort by abundance
SET HEADING ON
SORT TO TEMP1 ON ENTRY,NUMBER FOR D='E'.OR.D='H'.OR.D='O'.OR.D='X'.OR.D='I'
DO "COMPRESSION abundance.PRG"

CASE ANAL=4
* sort/interest
SET HEADING ON
IF CONDEN=1
SORT TO TEMP1 ON ENTRY,NUMBER FOR I>0
DO "COMPRESSION interest.PRG"
ELSE
SORT ON I/D,ENTRY TO TEMP1 FOR I>1
USE TEMP1
list off fields number,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,RFEND,INIT,I
CLOSE DATABASES
ERASE TEMP1.DBF
ENDIF

CASE ANAL=5
* arrange/location
SET HEADING ON
STORE 4 TO AMPLIFIER
? 'Nuclear:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Cytoplasmic:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Cytoskeleton:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Cell surface:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Intracellular membrane:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Mitochondrial:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF

```

```

? 'Secreted:'
SORT ON ENTRY, NUMBER FIELDS RFEND, NUMBER, L, D, F, Z, R, C, ENTRY, S, DESCRIPTOR, LENGTH, INIT, I, COMMENT
IF CONDENSE=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Other:'
SORT ON ENTRY, NUMBER FIELDS RFEND, NUMBER, L, D, F, Z, R, C, ENTRY, S, DESCRIPTOR, LENGTH, INIT, I, COMMENT
IF CONDENSE=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Unknown:'
SORT ON ENTRY, NUMBER FIELDS RFEND, NUMBER, L, D, F, Z, R, C, ENTRY, S, DESCRIPTOR, LENGTH, INIT, I, COMMENT
IF CONDENSE=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
IF CONDENSE=1
SET DEVICE TO PRINTER
SET PRINTER ON
EJECT
DO "Output heading.prg"
USE "Analysis location.dbf"
DO "Create bargraph.prg"
SET HEADING OFF
? '          FUNCTIONAL CLASS                TOTAL    UNIQUE    NEW    % TOTAL'
?
LIST OFF FIELDS Z, NAME, CLONES, GENES, NEW, PERCENT, GRAPH
CLOSE DATABASES
ERASE TEMP2.DBF
SET HEADING ON
*USE "SmartGuy:FoxBASE+/Mac:fox files:TEMPMASTER.dbf"
ENDIF

CASE ANAL=6
* arrange/distribution
SET HEADING ON
STORE 3 TO AMPLIFIER
? 'Cell/tissue specific distribution:'
SORT ON ENTRY, NUMBER FIELDS RFEND, NUMBER, L, D, F, Z, R, C, ENTRY, S, DESCRIPTOR, LENGTH, INIT, I, COMMENT
IF CONDENSE=1
DO "Compression distrib.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Non-specific distribution:'
SORT ON ENTRY, NUMBER FIELDS RFEND, NUMBER, L, D, F, Z, R, C, ENTRY, S, DESCRIPTOR, LENGTH, INIT, I, COMMENT
IF CONDENSE=1
DO "Compression distrib.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Unknown distribution:'
SORT ON ENTRY, NUMBER FIELDS RFEND, NUMBER, L, D, F, Z, R, C, ENTRY, S, DESCRIPTOR, LENGTH, INIT, I, COMMENT
IF CONDENSE=1
DO "Compression distrib.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
IF CONDENSE=1
SET DEVICE TO PRINTER
SET PRINTER ON

```

```

EJECT
DO "Output heading.prg"
USE "Analysis distribution.dbf"
DO "Create bargraph.prg"
SET HEADING OFF
? '          FUNCTIONAL CLASS          TOTAL  UNIQUE  $ TOTAL'
?
LIST OFF FIELDS P.NAME,CLONES,GENES,PERCENT,GRAPH
CLOSE DATABASES
ERASE TEMP2.DBF
SET HEADING ON
*USE "SmartGuy:FoxBASE+/Mac:fox files:TEMPMASTER.dbf"
ENDIF

CASE ANAL=7
* arrange/function
SET HEADING ON
STORE 10 TO AMPLIFIER
? '          BINDING PROTEINS'
?
? 'Surface molecules and receptors:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Calcium-binding proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Ligands and effectors:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Other binding proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
*EJECT
? '          ONCOGENES'
?
? 'General oncogenes:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'GTP-binding proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Viral elements:'

```

```

SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Kinases and Phosphatases:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Tumor-related antigens:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
*EJECT
? '
                                PROTEIN SYNTHETIC MACHINERY PROTEINS'
?
? 'Transcription and Nucleic Acid-binding proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Translation:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Ribosomal proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Protein processing:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
*EJECT
? '
                                ENZYMES'
?
? 'Ferropoteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Proteases and inhibitors:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE

```

```

DO "Normal subroutine 1"
ENDIF
? 'Oxidative phosphorylation:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Sugar metabolism:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Amino acid metabolism:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Nucleic acid metabolism:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Lipid metabolism:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Other enzymes:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
*EJECT
?
?
? 'Stress response:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Structural:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Other clones:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE

```

MISCELLANEOUS CATEGORIES'


```

DO "Normal subroutine 1"
ENDIF
? "Clones of unknown function:"
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF

IF CONDEN=1
EJECT
*SET DEVICE TO PRINTER
*SET PRINT ON
DO "Output heading.prg"
***
USE "Analysis function.dbf"
DO "Create bargraph.prg"
SET HEADING OFF
***
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",12 COLOR 0,0,0
***
? |
? |          FUNCTIONAL CLASS                      CLONES  GENES  TOTAL  TOTAL  NEW  DIST
? |
? |
***
*LIST OFF FIELDS P,NAME,CLONES,GENES,NEW,PERCENT,GRAPH,CCOMPANY
LIST OFF FIELDS P,NAME,CLONES,GENES,NEW,PERCENT,GRAPH
CLOSE DATABASES
ERASE TEMP2.DBF
SET HEADING ON
*USE "SmartGuy:FoxBASE+/Mac:fox files:TEMPMASTER.dbf"
ENDIF
CASE ANAL=8
DO "Subgroup summary 3.prg"
ENDCASE
DO "Test print.prg"
SET PRINT OFF
SET DEVICE TO SCREEN
CLOSE DATABASES
*ERASE TEMPLIB.DBF
*ERASE TEMPNUM.DBF
*ERASE TEMPDESIG.DBF
*ERASE SELECTED.DBF
CLEAR
LOOP
ENDDO

```

```

* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
    PACK
    COUNT TO UNIQUE
    COUNT TO NEWGENES FOR D='H'.OR.D='O'
    SW2=1
  LOOP
ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
  SKIP
  STORE ENTRY TO TESTB
  IF TESTA = TESTB
    DELETE
    DUP = DUP+1
  LOOP
ENDIF
GO MARK1.
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
GO TOP
STORE Z TO LOC
USE 'Analysis location.dbf'
LOCATE FOR Z=LOC
REPLACE CLONES WITH TOT
REPLACE GENES WITH UNIQUE
REPLACE NEW WITH NEWGENES
USE TEMP1
SORT ON RFEND/D TO TEMP2
USE TEMP2
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? '.clones'
? '
V Coincidence'
list off fields number, RFEND, L, D, F, Z, R, C, ENTRY, S, DESCRIPTOR, LENGTH, INIT, I

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE TEMPDESIG

```

```

* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
    PACK
    COUNT TO UNIQUE
    SW2=1
  LOOP
ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
  SKIP
  STORE ENTRY TO TESTB
  IF TESTA = TESTB
    DELETE
    DUP = DUP+1
  LOOP
ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
*BROWSE
*SET PRINTER ON
SORT ON DATE TO TEMP2
USE TEMP2
?? STR(UNIQUE,4,0)
?? ' genes, for a total of'
?? STR(TOT,4,0)
?? ' clones'
?
? ' V Coincidence'
COUNT TO P4 FOR I=4
IF P4>0
  ? STR(P4,3,0)
  ?? ' genes with priority = 4 (Secondary analysis:)'
  list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT for I=4
  ?
ENDIF
COUNT TO P3 FOR I=3
IF P3>0
  ? STR(P3,3,0)
  ?? ' genes with priority = 3 (Full insert sequence:)'
  list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT for I=3
  ?
ENDIF
COUNT TO P2 FOR I=2.
IF P2>0
  ? STR(P2,3,0)
  ?? ' genes with priority = 2 (Primary analysis complete:)'
  list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT for I=2
  ?
ENDIF
COUNT TO P1 FOR I=1
IF P1>0

```

```
? STR(P1,3,0)
?? ' genes with priority = 1 (Primary analysis needed:)'
list off fields number,RFEND,L,D,P,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT for I=1
ENDIF
```

```
*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE 'SmartGuy\FoxBASE+/Mac:fox files:clones.dbf'
```

```

* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
    PACK
    COUNT TO UNIQUE
    SW2=1
  LOOP
ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
  SKIP
  STORE ENTRY TO TESTB
  IF TESTA = TESTB
    DELETE
    DUP = DUP+1
  LOOP
ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
*BROWSE
*SET PRINTER ON
SORT ON NUMBER TO TEMP2
USE TEMP2

?? STR(UNIQUE,4,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? '
      V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE 'SmartGuy\FoxBASE+/Mac:fox files:clones.dbf'

```

* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS

USE TEMP1

COUNT TO TOT

REPLACE ALL RFEND WITH 1

MARK1 = 1

SW2=0

DO WHILE SW2=0 ROLL

IF MARK1 >= TOT

PACK

COUNT TO UNIQUE

COUNT TO NEWGENES FOR D='H'.OR.D='O'

SW2=1

LOOP

ENDIF

GO MARK1

DUP = 1

STORE ENTRY TO TESTA

SW = 0

DO WHILE SW=0 TEST

SKIP

STORE ENTRY TO TESTB

IF TESTA = TESTB

DELETE

DUP = DUP+1

LOOP

ENDIF

GO MARK1

REPLACE RFEND WITH DUP

MARK1 = MARK1+DUP

SW=1

LOOP

ENDDO TEST

LOOP

ENDDO ROLL

GO TOP

STORE R TO FUNC

USE "Analysis function.dbf"

LOCATE FOR P=FUNC

REPLACE CLONES WITH TOT

REPLACE GENES WITH UNIQUE

REPLACE NEW WITH NEWGENES.

USE TEMP1

SORT ON RFEND/D TO TEMP2

USE TEMP2

SET HEADING ON

?? STR(UNIQUE,5,0)

?? ' genes, for a total of '

?? STR(TOT,5,0)

?? ' clones'

? ' V Coincidence'

list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I

*SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",12 COLOR 0,0,

*list off fields RFEND,S,DESCRIPTOR

*SET PRINT OFF

CLOSE DATABASES

ERASE TEMP1.DBF

ERASE TEMP2.DBF

USE TEMPDESIG


```

* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
    PACK
    COUNT TO UNIQUE
    SW2=1
  LOOP
ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
  SKIP
  STORE ENTRY TO TESTB
  IF TESTA = TESTB
    DELETE
    DUP = DUP+1
  LOOP
ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
GO TOP
STORE F TO DIST
USE 'Analysis distribution.dbf'
LOCATE FOR P=DIST
REPLACE CLONES WITH TOT
REPLACE GENES WITH UNIQUE
USE TEMP1
sort on rfend/d to TEMP2
USE TEMP2
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? ' V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE TEMPDESIG

```

```

* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
    PACK
    COUNT TO UNIQUE
    SW2=1
  LOOP
ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
  SKIP
  STORE ENTRY TO TESTB
  IF TESTA = TESTB
    DELETE
    DUP = DUP+1
  LOOP
ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
GO TOP
USE TEMP1
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? '
? ' V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
USE TEMPDESIG

```

```

* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
COPY TO TEMP1 FOR
USE TEMP1
COUNT TO IDGENE FOR D='E'.OR.D='O'.OR.D='H'.OR.D='N'.OR.D='R'.OR.D='A'
DELETE FOR D='N'.OR.D='D'.OR.D='A'.OR.D='U'.OR.D='S'.OR.D='M'.OR.D='R'.OR.D='V'
PACK
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
    PACK
    COUNT TO UNIQUE
    SW2=1
    LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
  SKIP
  STORE ENTRY TO TESTB
  IF TESTA = TESTB
    DELETE
    DUP = DUP+1
    LOOP
  ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
*BROWSE
*SET PRINTER ON
SORT ON RFEND/D,NUMBER TO TEMP2
USE TEMP2
REPLACE ALL START WITH RFEND/IDGENE*10000
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? ' Coincidence V      V Clones/10000'
set heading off
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
list fields number,RFEND,START,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,INIT,I
*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"

```

```

* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO IDGENE FOR D='E'.OR.D='O'.OR.D='H'.OR.D='N'.OR.D='R'.OR.D='A'
DELETE FOR D='N'.OR.D='D'.OR.D='A'.OR.D='U'.OR.D='S'.OR.D='M'.OR.D='R'.OR.D='V'
PACK
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
    PACK
    COUNT TO UNIQUE
    SW2=1
  LOOP
ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
  SKIP
  STORE ENTRY TO TESTB
  IF TESTA = TESTB
    DELETE
    DUP = DUP+1
  LOOP
ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
*BROWSE
*SET PRINTER ON
SORT ON RFEND/D,NUMBER TO TEMP2
USE TEMP2
REPLACE ALL START WITH RFEND/IDGENE*10000
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? ' Coincidence V      V Clones/10000'
set heading off
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list fields number,RFEND,START,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,INIT,I
*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE 'SmartGuy:FoxBASE+/Mac:fox files:clones.dbf'

```

```
USE TEMP1
COUNT TO TOT
?? ' Total of'
?? STR(TOT,4,0)
?? ' clones'
?
*list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR,LENGTH,RFEND,INIT,I
list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR
CLOSE DATABASES
ERASE TEMP1.DBF
USE TEMPDESIG
```

```

*Lifescan menu; version 8-7-94
SET TALK OFF
set device to screen
CLEAR
USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
STORE LUPDATE() TO Update
GO BOTTOM
STORE RECNO() TO cloneno
STORE 6 TO Chooser
DO WHILE .T.
  * Program.: Lifeseq menu.fmt
  * Date....: 1/11/95
  * Version.: FoxBASE+/Mac, revision 1.10
  * Notes....: Format file Lifeseq menu
  *
  SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",268 COLOR 0,0,
  @ PIXELS 18,126 TO 77,365 STYLE 28479 COLOR 32767,-25600,-1,-16223,-16721,-15725
  @ PIXELS 110,29 TO 188,217 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
  @ PIXELS 45,161 SAY "LIFESEQ" STYLE 65536 FONT "Geneva",536 COLOR 0,0,-1,-1,7135,5884
  @ PIXELS 36,269 SAY "TM" STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,7135,5884
  @ PIXELS 63,143 SAY "Molecular Biology Desktop" STYLE 65536 FONT "Helvetica",18 COLOR 0,0,0,
  @ PIXELS 90,252 TO 251,467 STYLE 28447 COLOR 0,0,-1,-25600,-1,-1
  @ PIXELS 117,270 GET Chooser STYLE 65536 FONT "Chicago",12 PICTURE "@*RV Transcript profiles
  @ PIXELS 135,128 SAY Update STYLE 0 FONT "Geneva",12 SIZE 15,79 COLOR 0,0,0,-25600,-1,-1
  @ PIXELS 171,128 SAY cloneno STYLE 0 FONT "Geneva",12 SIZE 15,79 COLOR 0,0,0,-25600,-1,-1
  @ PIXELS 135,44 SAY "Last update:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,-1,-1
  @ PIXELS 171,44 SAY "Total clones:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,-1,-1
  @ PIXELS 45,296 SAY "v1.30" STYLE 65536 FONT "Geneva",782 COLOR 0,0,-1,-1,-1,-1
  *
  * EOF: Lifeseq menu.fmt
  READ
  DO CASE
  CASE Chooser=1
  DO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:Master analysis 3.prg"
  CASE Chooser=2
  DO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:Subtraction 2.prg"
  CASE Chooser=3
  DO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:Northern (single).prg"
  CASE Chooser=4
  USE "Libraries.dbf"
  BROWSE
  CASE Chooser=5
  DO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:See individual clone.prg"
  CASE Chooser=6
  DO "SmartGuy:FoxBASE+/Mac:fox files:Libraries:Output programs:Menu.prg"
  CASE Chooser=7
  CLEAR
  SCREEN 1 OFF
  RETURN
  ENDCASE

  LOOP
ENDDO

```



```

@1,30 SAY "Database Subset Analysis" STYLE 65536, FONT "Geneva", 274 COLOR 0,0,0,-1,-1,-1
?
?
?
? date()
?? '
?? TIME()
? 'Clone numbers '
?? STR(INITIATE,6,0)
?? ' through '
?? STR(TERMINATE,6,0)
? 'Libraries: '
IF ENTIRE=1
? 'All libraries'
ENDIF
IF ENTIRE=2
  MARK=1
  DO WHILE .T.
    IF MARK>STOPIT
      EXIT
    ENDIF
    USE SELECTED
    GO MARK
    ? ' '
    ?? TRIM(libname)
    STORE MARK+1 TO MARK
  LOOP
  ENDDO
ENDIF
? 'Designations: '
IF Ematch=0 .AND. Hmatch=0 .AND. Omatch=0
?? 'All'
ENDIF
IF Ematch=1
?? 'Exact,'
ENDIF
IF Hmatch=1
?? 'Human,'
ENDIF
IF Omatch=1
?? 'Other sp.'
ENDIF
IF CONDENSE=1
? 'Condensed format analysis'
ENDIF
IF ANAL=1
? 'Sorted by NUMBER'
ENDIF
IF ANAL=2
? 'Sorted by ENTRY'
ENDIF
IF ANAL=3
? 'Arranged by ABUNDANCE'
ENDIF
IF ANAL=4
? 'Sorted by INTEREST'
ENDIF
IF ANAL=5
? 'Arranged by LOCATION'
ENDIF
IF ANAL=6
? 'Arranged by DISTRIBUTION'
ENDIF
IF ANAL=7
? 'Arranged by FUNCTION'

```

```
ENDIF
? 'Total clones represented: '
?? STR(STARTOT,6,0)
? 'Total clones analyzed: '
?? STR(ANALTOT,6,0)
?
?
```

```
USE TEMP1
COUNT TO TOT
?? ' Total of'
?? STR(TOT,4,0)
?? ' clones'
?
*list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR,LENGTH,RFEND,INIT,I
list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR
CLOSE DATABASES
ERASE TEMP1.DBF
USE TEMPDESIG
```

```
USE TEMP1
COUNT TO TOT
?? ' Total of'
?? STR(TOT,4,0)
?? ' clones'
?
*list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR,LENGTH,RFEND,INIT,I
list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR
CLOSE DATABASES
ERASE TEMP1.DBF
USE TEMPDESIG
```

```

*Northern (single), version 11-25-94
close databases
SET TALK OFF
SET PRINT OFF
SET EXACT OFF
CLEAR
STORE ' ' TO Eobject
STORE ' ' TO Dobject
STORE 0 TO Numb
STORE 0 TO Zog
STORE 1 TO Bail
DO WHILE .T.
* Program.: Northern (single).fmt
* Date.....: 8/ 8/94
* Version.: FoxBASE+/Mac, revision 1.10
* Notes.....: Format file Northern (single)
*
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",12 COLOR 0,0,0
@ PIXELS 15,81 TO 46,397 STYLE 28447 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 89,79 TO 192,422 STYLE 28447 COLOR 0,0,0,-25600,-1,-1
@ PIXELS 115,98 SAY "Entry #:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 115,173 GET Eobject STYLE 0 FONT "Geneva",12 SIZE 15,142 COLOR 0,0,0,-1,-1,-1
@ PIXELS 145,89 SAY "Description" STYLE 65536 FONT "Geneva",12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 145,173 GET Dobject STYLE 0 FONT "Geneva",12 SIZE 15,241 COLOR 0,0,0,-1,-1,-1
@ PIXELS 35,89 SAY "Single Northern search screen" STYLE 65536 FONT "Geneva",274 COLOR 0,0,-
@ PIXELS 220,162 GET Bail STYLE 65536 FONT "Chicago",12 PICTURE "2*R Continue;Bail out" SIZE
@ PIXELS 175,98 SAY "Clone #:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 175,173 GET Numb STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,0,-1,-1,-1
@ PIXELS 80,152 SAY "Enter any ONE of the following:" STYLE 65536 FONT "Geneva",12 COLOR -1,
*
* EOF: Northern (single).fmt
READ
IF Bail=2
CLEAR
screen 1 off
RETURN
ENDIF
USE "SmartGuy:FoxBASE+/Mac:Fox files:Lookup.dbf"
SET TALK ON

IF Eobject<>'
STORE UPPER(Eobject) to Eobject
SET SAFETY OFF
SORT ON Entry TO "Lookup entry.dbf"
SET SAFETY ON
USE "Lookup entry.dbf"
LOCATE FOR Look=Eobject
IF .NOT.FOUND()
CLEAR
LOOP
ENDIF
BROWSE
STORE Entry TO Searchval
CLOSE DATABASES
ERASE "Lookup entry.dbf"
ENDIF

IF Dobject<>'
SET EXACT OFF
SET SAFETY OFF
SORT ON descriptor TO "Lookup descriptor.dbf"
SET SAFETY On
USE "Lookup descriptor.dbf"
LOCATE FOR UPPER(TRIM(descriptor))=UPPER(TRIM(Dobject))
IF .NOT.FOUND()
CLEAR

```

```

LOOP
ENDIF
BROWSE
STORE Entry TO Searchval
CLOSE DATABASES
ERASE "Lookup descriptor.dbf"
SET EXACT ON
ENDIF

IF Numb<>0
USE "SmartGuy:FoxBASE+/Mac:Fox files:clones.dbf"
GO Numb
BROWSE
STORE Entry TO Searchval
ENDIF

CLEAR
? 'Northern analysis for entry '
?? Searchval
?
? 'Enter Y to proceed'
WAIT TO OK
CLEAR
IF UPPER(OK)<>'Y'
screen 1 off
RETURN
ENDIF

* COMPRESSION SUBROUTINE FOR Library.dbf
? 'Compressing the Libraries file now...'
USE "SmartGuy:FoxBASE+/Mac:Fox files:libraries.dbf"
SET SAFETY OFF
SORT ON library TO "Compressed libraries.dbf"
* FOR entered>0
SET SAFETY ON
USE "Compressed libraries.dbf"
DELETE FOR entered=0
PACK
COUNT TO TOT
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
    IF MARK1 >= TOT
        PACK
        SW2=1
        LOOP
    ENDIF
GO MARK1
STORE library TO TESTA
SKIP
STORE Library TO TESTB
IF TESTA = TESTB
DELETE
ENDIF
MARK1 = MARK1+1
LOOP
ENDDO ROLL

* Northern analysis
CLEAR
? 'Doing the northern now...'
SET TALK ON
USE "SmartGuy:FoxBASE+/Mac:Fox files:clones.dbf"
SET SAFETY OFF
COPY TO "Hits.dbf" FOR entry=searchval
SET SAFETY ON

```



```
CLOSE DATABASES
SELECT 1
USE "Compressed libraries.dbf"
STORE RECCOUNT() TO Entries
SELECT 2
USE "Hits.dbf"
Mark=1
DO WHILE .T.
  SELECT 1
  IF Mark>Entries
    EXIT
  ENDIF
  GO MARK
  STORE library TO Jigger
  SELECT 2
  COUNT TO Zog FOR library=Jigger
  SELECT 1
  REPLACE hits with Zog
  Mark=Mark+1
  LOOP
ENDDO

SELECT 1
BROWSE FIELDS LIBRARY,LIBNAME,ENTERED,HITS AT 0,0
CLEAR
? 'Enter Y to print:'
WAIT TO PRINSET
IF UPPER(PRINSET)='Y'
  SET PRINT ON
  CLEAR
  EJECT.
  SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",14 COLOR 0,0,0
  ? 'DATABASE ENTRIES MATCHING ENTRY '
  ?? Searchval
  ? DATE()
  ?
  SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS FONT "Geneva",7 COLOR 0,0,0,
  LIST OFF FIELDS library,libname,entered,hits
  ?
  ?
  SELECT 2
  LIST OFF FIELDS NUMBER,LIBRARY,D,S,F,Z,R,ENTRY,DESCRIPTOR,RFSTART,START,RFEND
  SET TALK OFF
  SET PRINT OFF
  ENDIF
  CLOSE DATABASES
  SET TALK OFF
  CLEAR
  DO "Test print.prg"
  RETURN
```

TABLE 6

library	libname
ADENINB01	Inflamed adenoid
ADRENOR01	Adrenal gland (r)
ADRENOT01	Adrenal gland (T)
AMLBNOT01	AML blast cells (T)
BMARNOT01	Bone marrow
BMARNOT02	Bone marrow (T)
CARDNOT01	Cardiac muscle (T)
CHAONOT01	Chin. hamster ovary
CORNNOT01	Corneal stroma
FIBRAGT01	Fibroblast, AT 5
FIBRAGT02	Fibroblast, AT 30
FIBRANT01	Fibroblast, AT
FIBRNGT01	Fibroblast, uv 5
FIBRNGT02	Fibroblast, uv 30
FIBRNOT01	Fibroblast
FIBRNOT02	Fibroblast, normal
HMC1NOT01	Mast cell line HMC-1
HUVELPB01	HUVEC IFN,TNF,LPS
HUVENOB01	HUVEC control
HUVESTB01	HUVEC shear stress
HYPONOB01	Hypothalamus
KIDNNOT01	Kidney (T)
LIVRNOT01	Liver (T)
LUNGNOT01	Lung (T)
MUSCNOT01	Skeletal muscle (T)
OVIDNOB01	Oviduct
PANCN0T01	Pancreas, normal
PITUNOR01	Pituitary (r)
PITUNOT01	Pituitary (T)
PLACNOB01	Placenta
SINTNOT02	Small intestine (T)
SPLNFET01	Spleen+liver, fetal
SPLNNOT02	Spleen (T)
STOMNOT01	Stomach
SYNORAE01	Rheum. synovium
TBLYN0T01	T + B lymphoblast
TESTNOT01	Testis (T)
THP1NOB01	THP-1 control
THP1PEB01	THP phorbol
THP1PLB01	THP-1 phorbol LPS
U937NOT01	U937, monocytic leuk

number	library	d s f z r entry	descriptor	rfstart	rfend
2304	U937NOT01	E H C C T HUMEF1B	Elongation factor 1-beta	0	773
3240	HMC1NOT01	E H C C T HUMEF1B	Elongation factor 1-beta	0	773
3269	HMC1NOT01	E H C C T HUMEF1B	Elongation factor 1-beta	0	773
4693	HMC1NOT01	E H C C T HUMEF1B	Elongation factor 1-beta	0	773
8389	HMC1NOT01	E H C C T HUMEF1B	Elongation factor 1-beta	0	773
9139	HMC1NOT01	E H C C T HUMEF1B	Elongation factor 1-beta	0	773

WHAT IS CLAIMED IS:

1. A method of analyzing a specimen containing gene transcripts, said method comprising the steps of:
 - (a) producing a library of biological sequences;
 - 5 (b) generating a set of transcript sequences, where each of the transcript sequences in said set is indicative of a different one of the biological sequences of the library;
 - (c) processing the transcript sequences in a
10 programmed computer in which a database of reference transcript sequences indicative of reference biological sequences is stored, to generate an identified sequence value for each of the transcript sequences, where each said identified sequence value is indicative of a sequence
15 annotation and a degree of match between one of the transcript sequences and at least one of the reference transcript sequences; and
 - (d) processing each said identified sequence value to generate final data values indicative of a number of times
20 each identified sequence value is present in the library.
2. The method of claim 1, wherein step (a) includes the steps of:
 - obtaining a mixture of mRNA;
 - making cDNA copies of the mRNA;
 - 25 isolating a representative population of clones transfected with the cDNA and producing therefrom the library of biological sequences.
3. The method of claim 1, wherein the biological sequences are cDNA sequences.
- 30 4. The method of claim 1, wherein the biological sequences are RNA sequences.
5. The method of claim 1, wherein the biological sequences are protein sequences.

6. The method of claim 1, wherein a first value of said degree of match is indicative of an exact match, and a second value of said degree of match is indicative of a non-exact match.

- 5 7. A method of comparing two specimens containing gene transcripts, said method comprising:
- (a) analyzing a first specimen according to the method of claim 1;
 - (b) producing a second library of biological
10 sequences;
 - (c) generating a second set of transcript sequences, where each of the transcript sequences in said second set is indicative of a different one of the biological sequences of the second library;
 - 15 (d) processing the second set of transcript sequences in said programmed computer to generate a second set of identified sequence values known as further identified sequence values, where each of the further identified sequence values is indicative of a sequence annotation and
20 a degree of match between one of the biological sequences of the second library and at least one of the reference sequences;
 - (e) processing each said further identified sequence value to generate further final data values indicative of a
25 number of times each further identified sequence value is present in the second library; and
 - (f) processing the final data values from the first specimen and the further identified sequence values from the second specimen to generate ratios of transcript
30 sequences, each of said ratio values indicative of differences in numbers of gene transcripts between the two specimens.

8. A method of quantifying relative abundance of mRNA in a biological specimen, said method comprising the steps
35 of:

- (a) isolating a population of mRNA transcripts from the biological specimen;

(b) identifying genes from which the mRNA was transcribed by a sequence-specific method;

(c) determining numbers of mRNA transcripts corresponding to each of the genes; and

5 (d) using the mRNA transcript numbers to determine the relative abundance of mRNA transcripts within the population of mRNA transcripts.

9. A diagnostic method which comprises producing a gene transcript image, said method comprising the steps of:

10 (a) isolating a population of mRNA transcripts from a biological specimen;

(b) identifying genes from which the mRNA was transcribed by a sequence-specific method;

(c) determining numbers of mRNA transcripts
15 corresponding to each of the genes; and

(d) using the mRNA transcript numbers to determine the relative abundance of mRNA transcripts within the population of mRNA transcripts, where data determining the relative abundance values of mRNA transcripts is the gene
20 transcript image of the biological specimen.

10. The method of claim 9, further comprising:

(e) providing a set of standard normal and diseased gene transcript images; and

(f) comparing the gene transcript image of the
25 biological specimen with the gene transcript images of step (e) to identify at least one of the standard gene transcript images which most closely approximate the gene transcript image of the biological specimen.

11. The method of claim 9, wherein the biological
30 specimen is biopsy tissue, sputum, blood or urine.

12. A method of producing a gene transcript image, said method comprising the steps of

(a) obtaining a mixture of mRNA;

(b) making cDNA copies of the mRNA;

(c) inserting the cDNA into a suitable vector and using said vector to transfect suitable host strain cells which are plated out and permitted to grow into clones, each clone representing a unique mRNA;

5 (d) isolating a representative population of recombinant clones;

(e) identifying amplified cDNAs from each clone in the population by a sequence-specific method which identifies gene from which the unique mRNA was transcribed;

10 (f) determining a number of times each gene is represented within the population of clones as an indication of relative abundance; and

(g) listing the genes and their relative abundance in order of abundance, thereby producing the gene transcript
15 image.

13. The method of claim 12, also including the step of diagnosing disease by:

repeating steps (a) through (g) on biological specimens from random sample of normal and diseased humans,
20 encompassing a variety of diseases, to produce reference sets of normal and diseased gene transcript images;

obtaining a test specimen from a human, and producing a test gene transcript image by performing steps (a) through (g) on said test specimen;

25 comparing the test gene transcript image with the reference sets of gene transcript images; and

identifying at least one of the reference gene transcript images which most closely approximates the test gene transcript image.

30 14. A computer system for analyzing a library of biological sequences, said system including:

means for receiving a set of transcript sequences, where each of the transcript sequences is indicative of a different one of the biological sequences of the library;
35 and

means for processing the transcript sequences in the computer system in which a database of reference transcript

sequences indicative of reference biological sequences is stored, wherein the computer is programmed with software for generating an identified sequence value for each of the transcript sequences, where each said identified sequence value is indicative of a sequence annotation and a degree of match between a different one of the biological sequences of the library and at least one of the reference transcript sequences, and for processing each said identified sequence value to generate final data values indicative of a number of times each identified sequence value is present in the library.

15. The system of claim 14, also including:

library generation means for producing the library of biological sequences and generating said set of transcript sequences from said library.

16. The system of claim 15, wherein the library generation means includes:

means for obtaining a mixture of mRNA;
means for making cDNA copies of the mRNA;
20 means for inserting the cDNA copies into cells and permitting the cells to grow into clones;
means for isolating a representative population of the clones and producing therefrom the library of biological sequences.

SYBASE database Structure

Library Preparation

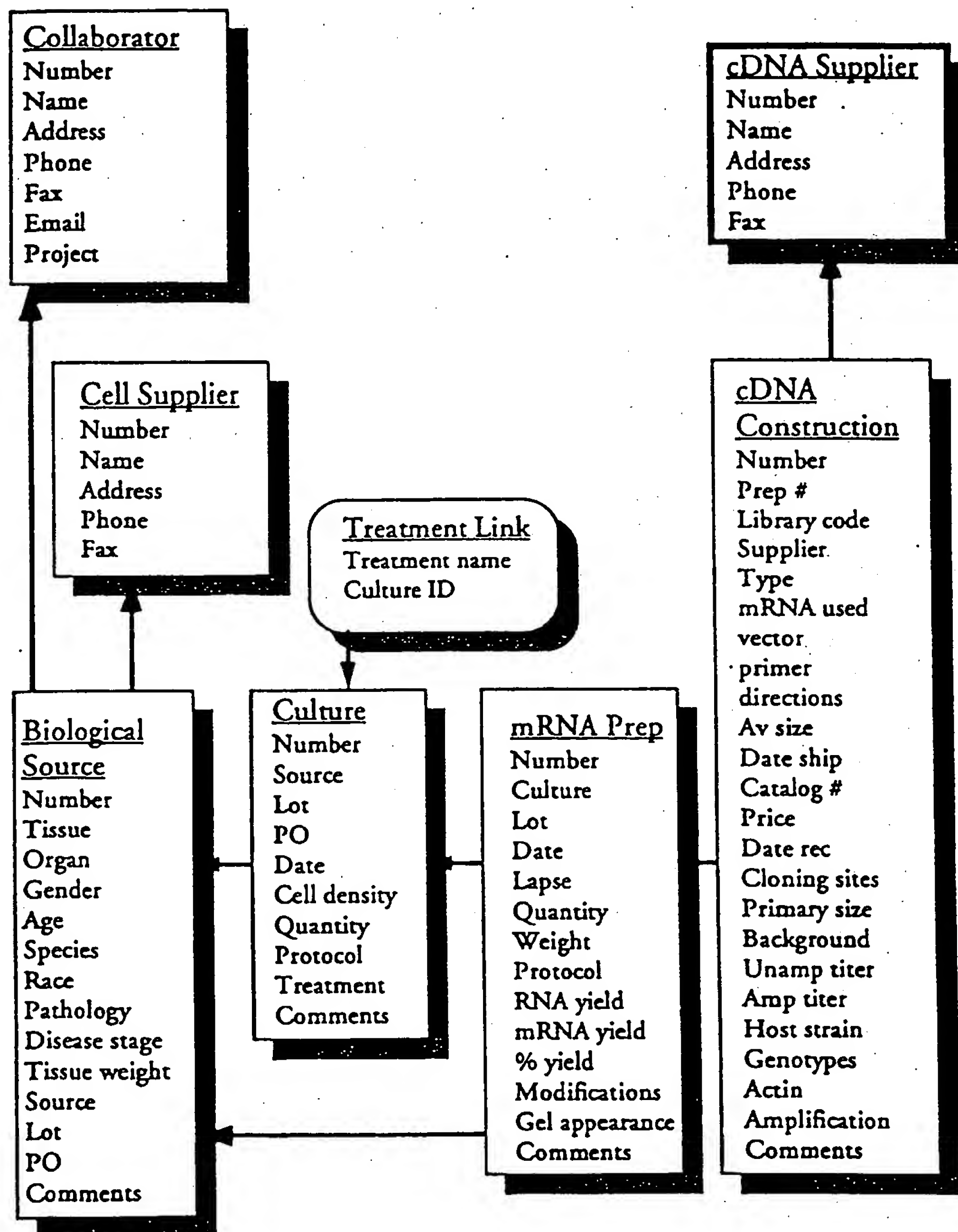


Figure 1

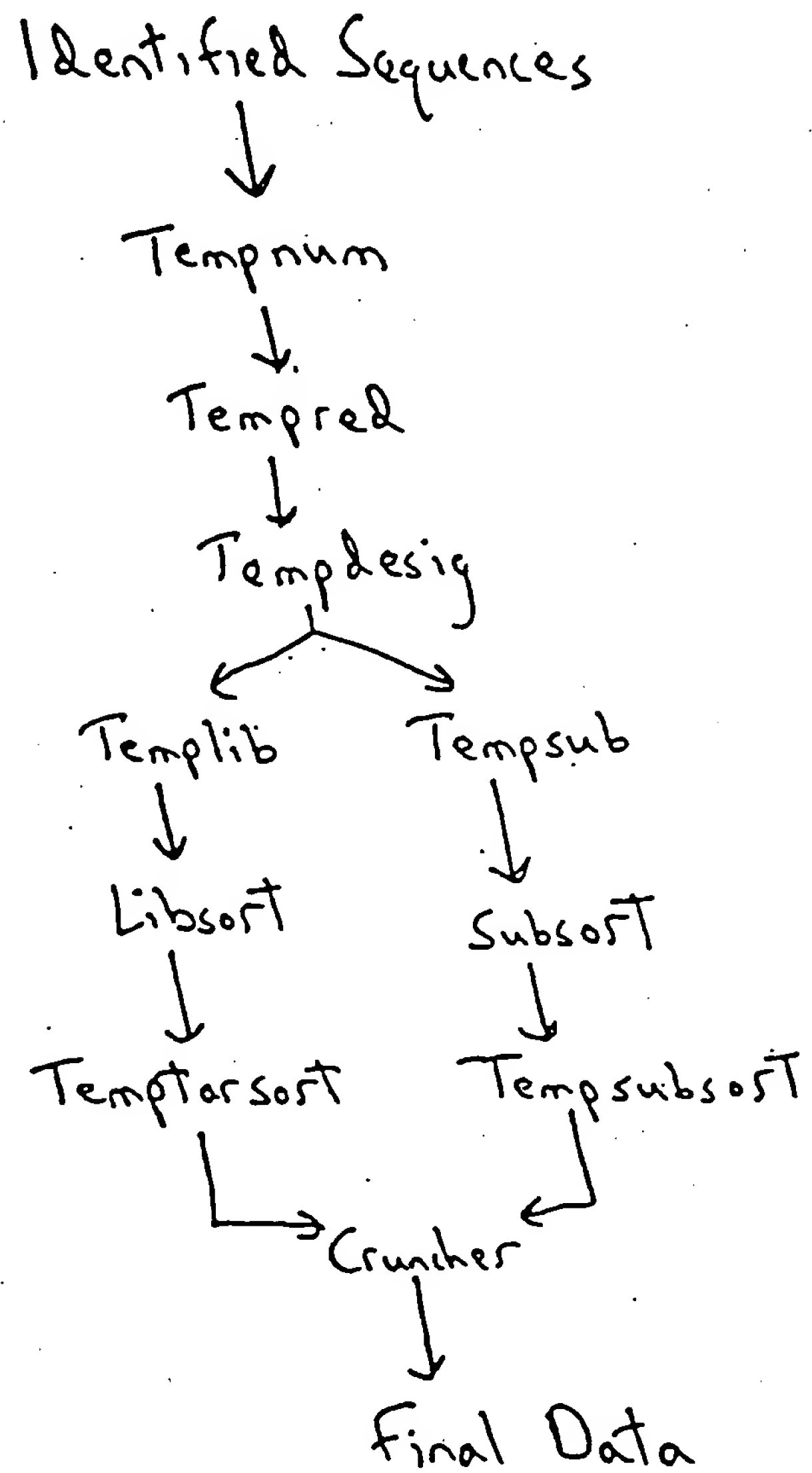


Figure 2

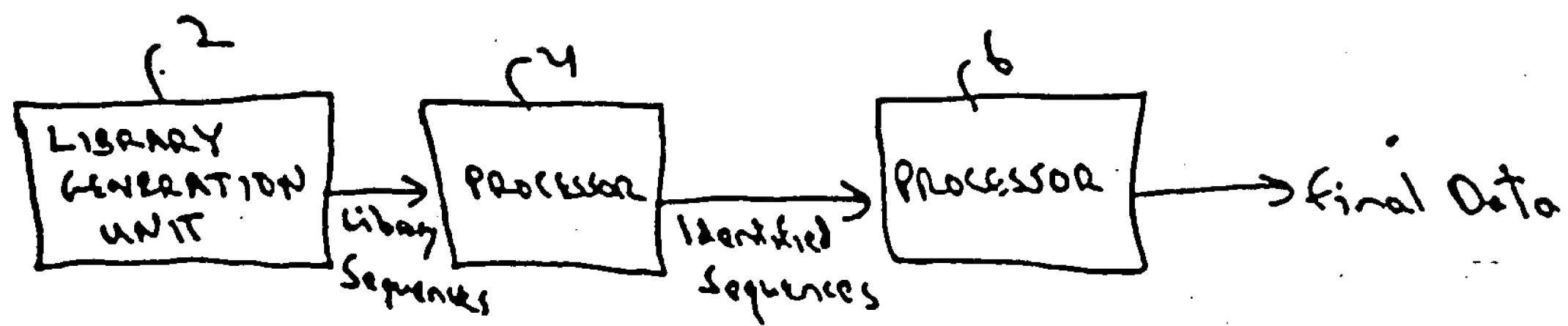


Figure 3

Incyte Bioinformatics Process

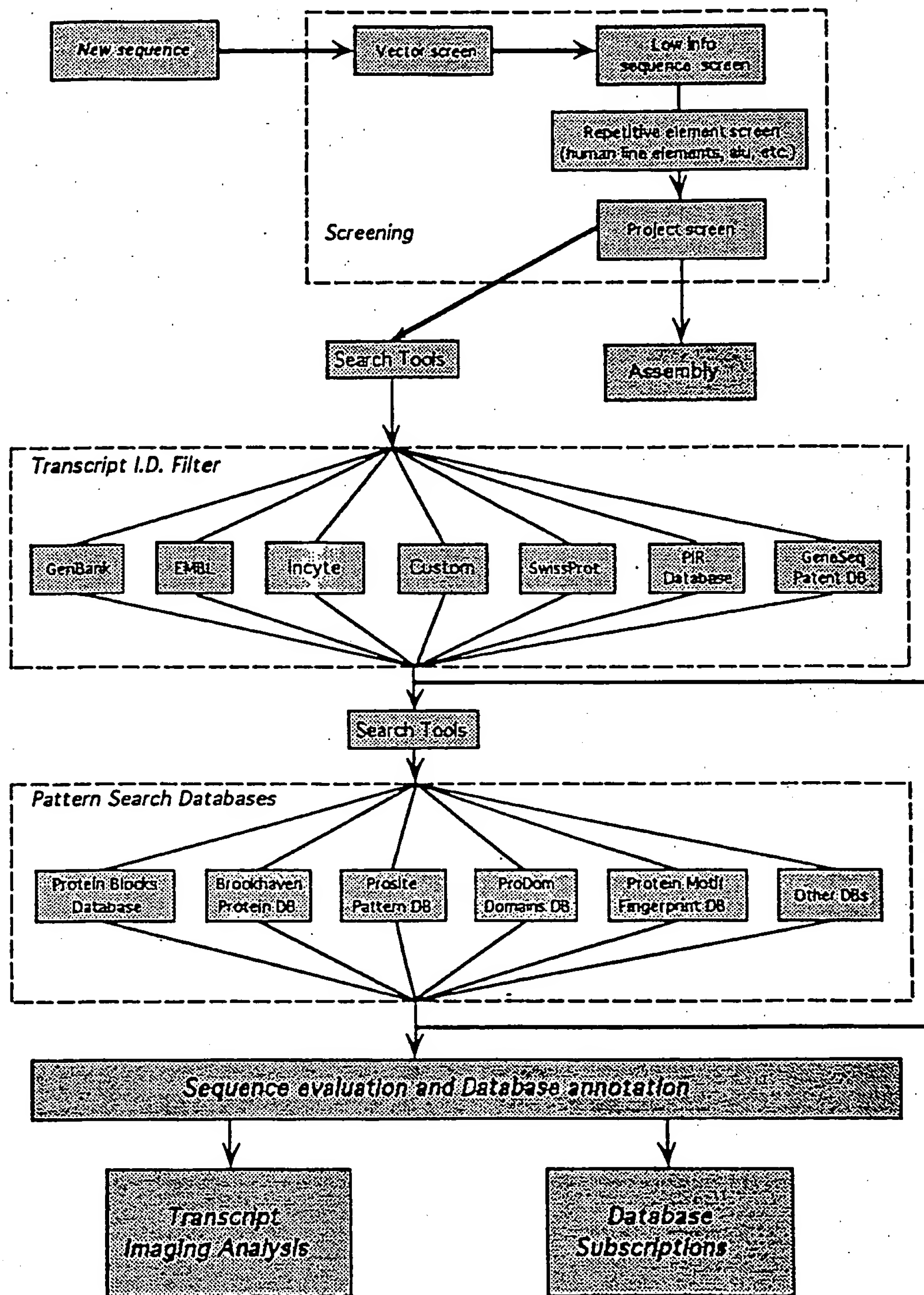


Figure 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US95/01160

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) : C12Q 1/68; G06F 15/00

US CL : 435/6; 364/413.02

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6; 364/413.02

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CAS ONLINE, APS, transcript, transcripts, cdan#, mrna#, frequenc?, distribut?, abundanc?

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	IntelliGenetics Suite, Release 5.4, Advanced Training Manual, issued January 1993 by IntelliGenetics, Inc. 700 East El Camino Real, Mountain View, California 94040, United States of America, pages (1-6)-(1-19) and (2-9)-(2-14), see entire document.	15 and 16
---		-----
Y		1-14
Y	Science, Volume 252, issued 21 June 1991, M.D. Adams et al, "Complementary DNA sequencing: Expressed sequence tags and human genome project", pages 1651-1656, see entire document.	1-16



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:	*T	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
*A		document defining the general state of the art which is not considered to be of particular relevance
*E		earlier document published on or after the international filing date
*L		document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
*O		document referring to an oral disclosure, use, exhibition or other means
*P		document published prior to the international filing date but later than the priority date claimed
	*X	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
	*Y	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
	*A	document member of the same patent family

Date of the actual completion of the international search

27 APRIL 1995

Date of mailing of the international search report

04 MAY 1995

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer

JAMES MARTINELL

Telephone No. (703) 308-0196

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US95/01160

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	Nucleic Acids Research, Volume 19, No. 25, issued 1991, E. Hara et al, "Subtractive cDNA cloning using oligo(dT) ₃₀ -latex and PCR: isolation of cDNA clones specific to undifferentiated human embryonal carcinoma cells", pages 7097-7104, see entire document.	1-16
X — Y	Nature Genetics, Volume 2, No. 3, issued November 1992, K. Okubo et al, "Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression", pages 173-179, see narrative text portion of entire document.	1, 3 ----- 2 and 4-16

Axl1p sequence following Ser²⁰⁰ and occurs within the domain of Axl1p that shows homology with hIDE (14). To delete the complete STE23 sequence and create the *ste23Δ::URA3* mutation, polymerase chain reaction (PCR) primers (5'-TCGGAAGACCTCAT-TCTTGCTCATTTTGATATTGCTC-3' and 5'-GCTACAAACAGC-GTCGACTTGAATGCCCGACATCTTGGACTGT-GGGGTATTTCACACCG-3') were used to amplify the URA3 sequence of pRS316, and the reaction product was transformed into yeast for one-step gene replacement [R. Rothstein, *Methods Enzymol.* 194, 281 (1991)]. To create the *axl1Δ::LEU2* mutation contained on p114, a 5.0-kb *Sal*I fragment from pAXL1 was cloned into pUC19, and an internal 4.0-kb *Hpa*I-*Xho*I fragment was replaced with a *LEU2* fragment. To construct the *ste23Δ::LEU2* allele (a deletion corresponding to 931 amino acids) carried on p153, a *LEU2* fragment was used to replace the 2.8-kb *Pml*I-Ecl136 II fragment of STE23, which occurs within a 6.2-kb *Hind*III-BglII genomic fragment carried on pSP72 (Promega). To create YEpMFA1, a 1.8-kb *Bam*HI fragment containing MFA1, from pKK16 [K. Kuchler, R. E. Sterne, J. Thormer, *EMBO J.* 8, 3973 (1989)], was ligated into the *Bam*HI site of YEp351 [J. E. Hill, A. M. Myers, T. J. Koerner, A. Tzagoloff, *Yeast* 2, 163 (1986)].

24. J. Chant and I. Herskowitz, *Cell* 65, 1203 (1991).
25. B. W. Matthews, *Acc. Chem. Res.* 21, 333 (1988).
26. K. Kuchler, H. G. Dohman, J. Thormer, *J. Cell Biol.* 120, 1203 (1993); R. Kolling and C. P. Hollenberg, *EMBO J.* 13, 3281 (1994); C. Berkower, D. Loeryza, S. Michaels, *Mol. Biol. Cell* 5, 1185 (1994).
27. A. Bender and J. R. Pringle, *Proc. Natl. Acad. Sci. U.S.A.* 86, 9976 (1989); J. Chant, K. Corrado, J. R. Pringle, I. Herskowitz, *Cell* 65, 1213 (1991); S. Powers, E. Gonzales, T. Christensen, J. Cubert, D. Broek, *ibid.*, p. 1225; H. O. Park, J. Chant, I. Herskowitz, *Nature* 365, 269 (1993); J. Chant, *Trends Genet.* 10, 328 (1994); _____ and J. R. Pringle, *J. Cell Biol.* 129, 751 (1995); J. Chant, M. Mischke, E. Mitchell, I. Herskowitz, J. R. Pringle, *ibid.*, p. 767.
28. G. F. Sprague Jr., *Methods. Enzymol.* 194, 77 (1991).
29. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
30. A W303 1A derivative, SY2625 (*MATa ura3-1 leu2-3, 112 trp1-1 ade2-1 can1-100 ssf1Δ mfa2Δ::FUS1-lacZ his3Δ::FUS1-HIS3*), was the parent strain for the mutant search. SY2625 derivatives for the mating assays, secreted pheromone assays, and the pulse-chase experiments included the following strains: Y49 (*ste22-1*), Y115 (*mfa1Δ::LEU2*), Y142 (*axl1Δ::URA3*), Y173 (*axl1Δ::LEU2*), Y220 (*axl1Δ::URA3 ste23Δ::URA3*), Y221 (*ste23Δ::URA3*), Y231 (*axl1Δ::LEU2 ste23Δ::LEU2*), and Y233 (*ste23Δ::LEU2*). *MATa* derivatives of SY2625 included the following strains: Y199 (SY2625 made *MATa*), Y278 (*ste22-1*), Y195 (*mfa1Δ::LEU2*), Y196 (*axl1Δ::LEU2*), and Y197 (*axl1Δ::URA3*). The EG123 (*MATa leu2 ura3 trp1 can1 his4*) genetic background was used to create a set of strains for analysis of bud site selection. EG123 derivatives included the following strains: Y175 (*axl1Δ::LEU2*), Y223 (*axl1Δ::URA3*), Y234 (*ste23Δ::LEU2*), and Y272 (*axl1Δ::LEU2 ste23Δ::LEU2*). *MATa* derivatives of EG123 included the following strains: Y214 (EG123 made *MATa*) and Y293 (*axl1Δ::LEU2*). All strains were generated by means of standard genetic or molecular methods involving the appropriate constructs (23). In particular, the *axl1 ste23* double mutant strains were created by crossing of the appropriate *MATa ste23* and *MATa axl1* mutants, followed by sporulation of the resultant diploid and isolation of the double mutant from nonparental di-type tetrads. Gene disruptions were confirmed with either PCR or Southern (DNA) analysis.
31. p129 is a YEp352 [J. E. Hill, A. M. Myers, T. J. Koerner, A. Tzagoloff, *Yeast* 2, 163 (1986)] plasmid containing a 5.5-kb *Sal*I fragment of pAXL1. p151 was derived from p129 by insertion of a linker at the *Bgl*II site within AXL1, which led to an in-frame insertion of the hemagglutinin (HA) epitope (DQYPTDVPDYA) (29) between amino acids 854 and 855 of the AXL1 prod-

uct. pC225 is a KS+ (Stratagene) plasmid containing a 0.5-kb *Bam*HI-SstI fragment from pAXL1. Substitution mutations of the proposed active site of Axl1p were created with the use of pC225 and site-specific mutagenesis involving appropriate synthetic oligonucleotides (*axl1-H68A*, 5'-GTGCTCACAAGGCT-GCCAAACCGGC-3'; *axl1-E71A*, 5'-AAGAATCAT-GTGCGCACAAGGTGGCG-3'; and *axl1-E71D*, 5'-AAGAATCATGTGATCACAAGGTGGCG-3'). The mutations were confirmed by sequence analysis. After mutagenesis, the 0.4-kb *Bam*HI-MscI fragment from the mutagenized pC225 plasmids was transferred into pAXL1 to create a set of pRS316 plasmids carrying different AXL1 alleles, p124 (*axl1-H68A*), p130 (*axl1-E71A*), and p132 (*axl1-E71D*). Similarly, a set of HA-tagged alleles carried on YEp352 were created after replacement of the p151 *Bam*HI-MscI fragment, to generate p161 (*axl1-E71A*), p162 (*axl1-*

32

N. Davis, T. Favaro, C. de Hoog, and S. Kim for comments on the manuscript. Supported by a grant to C.B. from the Natural Sciences and Engineering Research Council of Canada. Support for M.N.A. was from a California Tobacco-Related Disease Research Program postdoctoral fellowship (4FT-0083).

22 June 1995; accepted 21 August 1995

Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray

Mark Schena,* Dari Shalon,*† Ronald W. Davis, Patrick O. Brown‡

A high-capacity system was developed to monitor the expression of many genes in parallel. Microarrays prepared by high-speed robotic printing of complementary DNAs on glass were used for quantitative expression measurements of the corresponding genes. Because of the small format and high density of the arrays, hybridization volumes of 2 microliters could be used that enabled detection of rare transcripts in probe mixtures derived from 2 micrograms of total cellular messenger RNA. Differential expression measurements of 45 *Arabidopsis* genes were made by means of simultaneous, two-color fluorescence hybridization.

The temporal, developmental, topographical, histological, and physiological patterns in which a gene is expressed provide clues to its biological role. The large and expanding database of complementary DNA (cDNA) sequences from many organisms (1) presents the opportunity of defining these patterns at the level of the whole genome.

For these studies, we used the small flowering plant *Arabidopsis thaliana* as a model organism. *Arabidopsis* possesses many advantages for gene expression analysis, including the fact that it has the smallest genome of any higher eukaryote examined to date (2). Forty-five cloned *Arabidopsis* cDNAs (Table 1), including 14 complete sequences and 31 expressed sequence tags (ESTs), were used as gene-specific targets. We obtained the ESTs by selecting cDNA clones at random from an *Arabidopsis* cDNA library. Sequence analysis revealed that 28 of the 31 ESTs matched sequences

in the database (Table 1). Three additional cDNAs from other organisms served as controls in the experiments.

The 48 cDNAs, averaging ~1.0 kb, were amplified with the polymerase chain reaction (PCR) and deposited into individual wells of a 96-well microtiter plate. Each sample was duplicated in two adjacent wells to allow the reproducibility of the arraying and hybridization process to be tested. Samples from the microtiter plate were printed onto glass microscope slides in an area measuring 3.5 mm by 5.5 mm with the use of a high-speed arraying machine (3). The arrays were processed by chemical and heat treatment to attach the DNA sequences to the glass surface and denature them (3). Three arrays, printed in a single lot, were used for the experiments here. A single microtiter plate of PCR products provides sufficient material to print at least 500 arrays.

Fluorescent probes were prepared from total *Arabidopsis* mRNA (4) by a single round of reverse transcription (5). The *Arabidopsis* mRNA was supplemented with human acetylcholine receptor (AChR) mRNA at a dilution of 1:10,000 (w/w) before cDNA synthesis, to provide an internal standard for calibration (5). The resulting fluorescently labeled cDNA mixture was hybridized to an array at high stringency (6) and scanned

M. Schena and R. W. Davis, Department of Biochemistry, Beckman Center, Stanford University Medical Center, Stanford, CA 94305, USA.

D. Shalon and P. O. Brown, Department of Biochemistry and Howard Hughes Medical Institute, Beckman Center, Stanford University Medical Center, Stanford, CA 94305, USA.

*These authors contributed equally to this work.

†Present address: Syntex, Palo Alto, CA 94303, USA.

‡To whom correspondence should be addressed. E-mail: pbrown@cmgm.stanford.edu

with a laser (3). A high-sensitivity scan gave signals that saturated the detector at nearly all of the *Arabidopsis* target sites (Fig. 1A). Calibration relative to the AChR mRNA standard (Fig. 1A) established a sensitivity limit of $\sim 1:50,000$. No detectable hybridization was observed to either the rat glucocorticoid receptor (Fig. 1A) or the yeast TRP4 (Fig. 1A) targets even at the highest scanning sensitivity. A moderate-sensitivity scan

of the same array allowed linear detection of the more abundant transcripts (Fig. 1B). Quantitation of both scans revealed a range of expression levels spanning three orders of magnitude for the 45 genes tested (Table 2). RNA blots (7) for several genes (Fig. 2) corroborated the expression levels measured with the microarray to within a factor of 5 (Table 2).

Differential gene expression was investi-

gated with a simultaneous, two-color hybridization scheme, which served to minimize experimental variation inherent in the comparison of independent hybridizations. Fluorescent probes were prepared from two mRNA sources with the use of reverse transcriptase in the presence of fluorescein- and lissamine-labeled nucleotide analogs, respectively (5). The two probes were then mixed together in equal proportions, hybridized to a single array, and scanned separately for fluorescein and lissamine emission after independent excitation of the two fluorophores (3).

To test whether overexpression of a single gene could be detected in a pool of total *Arabidopsis* mRNA, we used a microarray to analyze a transgenic line overexpressing the single transcription factor HAT4 (8). Fluorescent probes representing mRNA from wild-type and HAT4-transgenic plants were labeled with fluorescein and lissamine, respectively; the two probes were then mixed and hybridized to a single array. An intense hybridization signal was observed at the position of the HAT4 cDNA in the lissamine-specific scan (Fig. 1D), but not in the fluorescein-specific scan of the same array (Fig. 1C). Calibration with AChR mRNA added to the fluorescein and lissamine cDNA synthesis reactions at dilutions of 1:10,000 (Fig. 1C) and 1:100 (Fig. 1D), respectively, revealed a 50-fold elevation of HAT4 mRNA in the transgenic line relative to its abundance in wild-type plants (Table 2). This magnitude of HAT4 overexpression matched that inferred from the Northern (RNA) analysis within a factor of 2 (Fig. 2 and Table 2). Expression of all the other genes monitored on the array differed by less than a factor of 5 between HAT4-transgenic and wild-type plants (Fig. 1, C

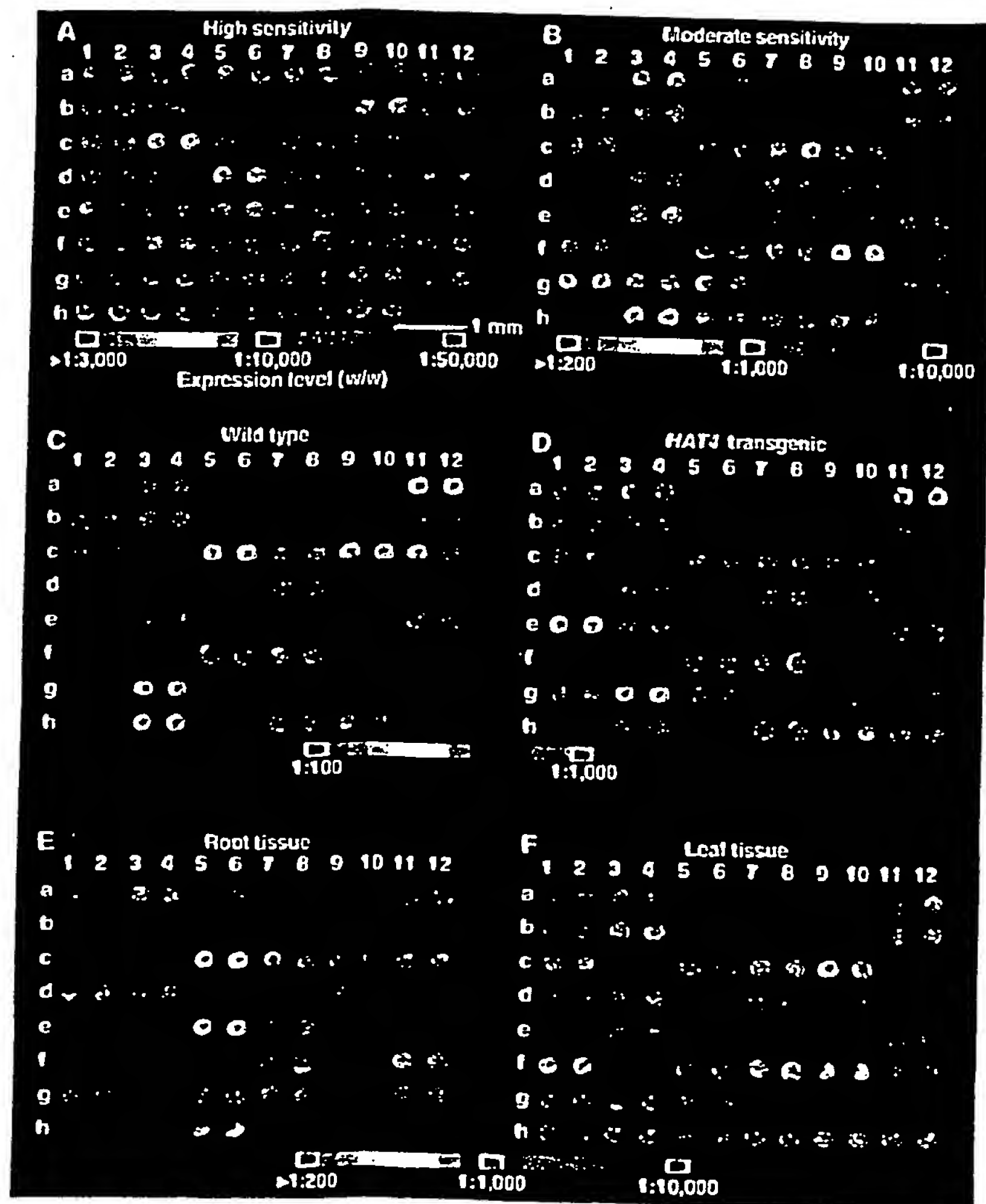


Fig. 1. Gene expression monitored with the use of cDNA microarrays. Fluorescent scans represented in pseudocolor correspond to hybridization intensities. Color bars were calibrated from the signal obtained with the use of known concentrations of human AChR mRNA in independent experiments. Numbers and letters on the axes mark the position of each cDNA. (A) High-sensitivity fluorescein scan after hybridization with fluorescein-labeled cDNA derived from wild-type plants. (B) Same array as in (A) but scanned at moderate sensitivity. (C and D) A single array was probed with a 1:1 mixture of fluorescein-labeled cDNA from wild-type plants and lissamine-labeled cDNA from HAT4-transgenic plants. The single array was then scanned successively to detect the fluorescein fluorescence corresponding to mRNA from wild-type plants (C) and the lissamine fluorescence corresponding to mRNA from HAT4-transgenic plants (D). (E and F) A single array was probed with a 1:1 mixture of fluorescein-labeled cDNA from root tissue and lissamine-labeled cDNA from leaf tissue. The single array was then scanned successively to detect the fluorescein fluorescence corresponding to mRNAs expressed in roots (E) and the lissamine fluorescence corresponding to mRNAs expressed in leaves (F).

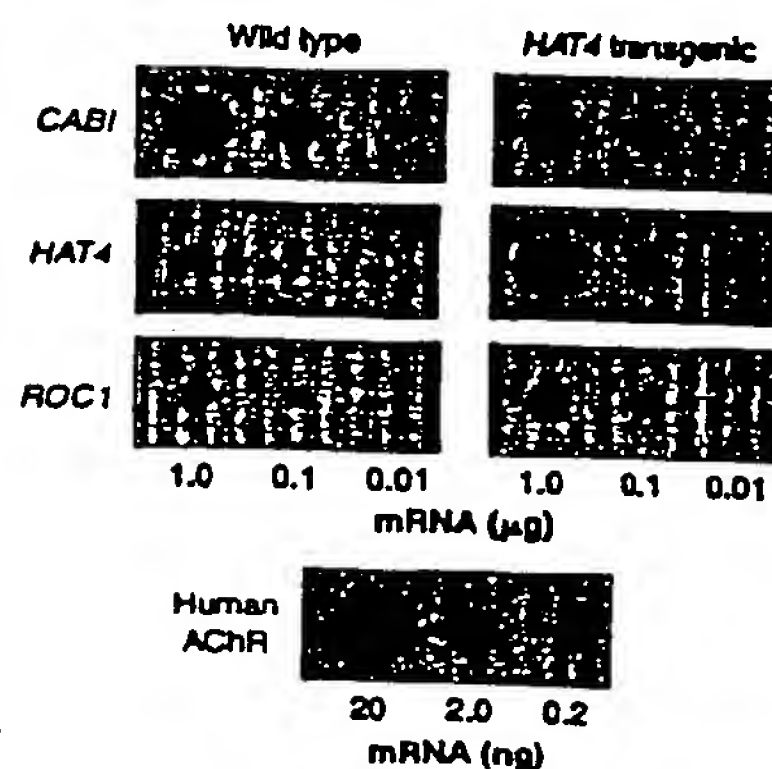


Fig. 2. Gene expression monitored with RNA (Northern) blot analysis. Designated amounts of mRNA from wild-type and HAT4-transgenic plants were spotted onto nylon membranes and probed with the cDNAs indicated. Purified human AChR mRNA was used for calibration.

and D, and Table 2). Hybridization of fluorescein-labeled glucocorticoid receptor cDNA (Fig. 1C) and lissamine-labeled TRP4 cDNA (Fig. 1D) verified the presence of the negative control targets and the lack of optical cross talk between the two fluorophores.

To explore a more complex alteration in expression patterns, we performed a second two-color hybridization experiment with fluorescein- and lissamine-labeled probes prepared from root and leaf mRNA, respectively. The scanning sensitivities for the two fluorophores were normalized by matching the signals resulting from AChR

mRNA, which was added to both cDNA synthesis reactions at a dilution of 1:1000 (Fig. 1, E and F). A comparison of the scans revealed widespread differences in gene expression between root and leaf tissue (Fig. 1, E and F). The mRNA from the light-regulated CABI gene was ~500-fold more abundant in leaf (Fig. 1F) than in root tissue (Fig. 1E). The expression of 26 other genes differed between root and leaf tissue by more than a factor of 5 (Fig. 1, E and F).

The HAT4-transgenic line we examined has elongated hypocotyls, early flowering, poor germination, and altered pigmentation (8). Although changes in expression were

observed for HAT4, large changes in expression were not observed for any of the other 44 genes we examined. This was somewhat surprising, particularly because comparative analysis of leaf and root tissue identified 27 differentially expressed genes. Analysis of an expanded set of genes may be required to identify genes whose expression changes upon HAT4 overexpression; alternatively, a comparison of mRNA populations from specific tissues of wild-type and HAT4-transgenic plants may allow identification of downstream genes.

At the current density of robotic printing, it is feasible to scale up the fabrication process to produce arrays containing 20,000 cDNA targets. At this density, a single array would be sufficient to provide gene-specific targets encompassing nearly the entire repertoire of expressed genes in the *Arabidopsis* genome (2). The availability of 20,274 ESTs from *Arabidopsis* (1, 9) would provide a rich source of templates for such studies.

The estimated 100,000 genes in the human genome (10) exceeds the number of *Arabidopsis* genes by a factor of 5 (2). This modest increase in complexity suggests that similar cDNA microarrays, prepared from the rapidly growing repertoire of human ESTs (1), could be used to determine the expression patterns of tens of thousands of human genes in diverse cell types. Coupling an amplification strategy to the reverse transcription reaction (11) could make it feasible to monitor expression even in minute tissue samples. A wide variety of acute and chronic physiological and pathological conditions might lead to characteristic changes in the patterns of gene expression in peripheral blood cells or other easily sampled tissues. In concert with cDNA microarrays for monitoring complex expression patterns, these tissues might therefore serve as sensitive *in vivo* sensors for clinical diagnosis. Microarrays of cDNAs could thus provide a useful link between human gene sequences and clinical medicine.

Table 2. Gene expression monitoring by microarray and RNA blot analyses; tg, HAT4-transgenic. See Table 1 for additional gene information. Expression levels (w/w) were calibrated with the use of known amounts of human AChR mRNA. Values for the microarray were determined from microarray scans (Fig. 1); values for the RNA blot were determined from RNA blots (Fig. 2).

Gene	Expression level (w/w)	
	Microarray	RNA blot
CABI	1:48	1:83
CABI (tg)	1:120	1:150
HAT4	1:8300	1:6300
HAT4 (tg)	1:150	1:210
ROC1	1:1200	1:1800
ROC1 (tg)	1:260	1:1300

Table 1. Sequences contained on the cDNA microarray. Shown is the position, the known or putative function, and the accession number of each cDNA in the microarray (Fig. 1). All but three of the ESTs used in this study matched a sequence in the database. NADH, reduced form of nicotinamide adenine dinucleotide; ATPase, adenosine triphosphatase; GTP, guanosine triphosphate.

Position	cDNA	Function	Accession number
a1, 2	AChR	Human AChR	-
a3, 4	EST3	Actin	H36236
a5, 6	EST6	NADH dehydrogenase	Z27010
a7, 8	AAC1	Actin 1	M20016
a9, 10	EST12	Unknown	U36594†
a11, 12	EST13	Actin	T45783
b1, 2	CABI	Chlorophyll a/b binding	M85150
b3, 4	EST17	Phosphoglycerate kinase	T44490
b5, 6	G44	Gibberellin acid biosynthesis	L37126
b7, 8	EST19	Unknown	U36595†
b9, 10	GBF-1	G-box binding factor 1	X63894
b11, 12	EST23	Elongation factor	X52256
c1, 2	EST29	Aldolase	T04477
c3, 4	GBF-2	G-box binding factor 2	X63895
c5, 6	EST34	Chloroplast protease	R87034
c7, 8	EST35	Unknown	T14152
c9, 10	EST41	Catalase	T22720
c11, 12	rGR	Rat glucocorticoid receptor	M14053
d1, 2	EST42	Unknown	U36596†
d3, 4	EST45	ATPase	J04185
d5, 6	HAT1	Homeobox-leucine zipper 1	U09332
d7, 8	EST46	Light harvesting complex	T04063
d9, 10	EST49	Unknown	T76267
d11, 12	HAT2	Homeobox-leucine zipper 2	U09335
e1, 2	HAT4	Homeobox-leucine zipper 4	M90394
e3, 4	EST50	Phosphoribulokinase	T04344
e5, 6	HAT5	Homeobox-leucine zipper 5	M90416
e7, 8	EST51	Unknown	Z33675
e9, 10	HAT22	Homeobox-leucine zipper 22	U09336
e11, 12	EST52	Oxygen evolving	T21749
f1, 2	EST59	Unknown	Z34607
f3, 4	KNAT1	Knotted-like homeobox 1	U14174
f5, 6	EST60	RuBisCO small subunit	X14564
f7, 8	EST69	Translation elongation factor	T42799
f9, 10	PPH1	Protein phosphatase 1	U34803
f11, 12	EST70	Unknown	T44621
g1, 2	EST75	Chloroplast protease	T43698
g3, 4	EST78	Unknown	R65481
g5, 6	ROC1	Cyclophilin	L14844
g7, 8	EST82	GTP binding	X59152
g9, 10	EST83	Unknown	Z33795
g11, 12	EST84	Unknown	T45278
h1, 2	EST91	Unknown	T13832
h3, 4	EST96	Unknown	R64816
h5, 6	SAR1	Synaptobrevin	M90418
h7, 8	EST100	Light harvesting complex	Z18205
h9, 10	EST103	Light harvesting complex	X03909
h11, 12	TRP4	Yeast tryptophan biosynthesis	X04273

*Proprietary sequence of Stratagene (La Jolla, California).

†No match in the database; novel EST.

REFERENCES AND NOTES

1. The current EST database (dbEST release 091495) from the National Center for Biotechnology Information (Bethesda, MD) contains a total of 322,225 entries, including 255,845 from the human genome and 21,044 from Arabidopsis. Access is available via the World Wide Web (<http://www.ncbi.nlm.nih.gov>).
2. E. M. Meyerowitz and R. E. Pruitt, *Science* 229, 1214 (1985); R. E. Pruitt and E. M. Meyerowitz, *J. Mol. Biol.* 187, 169 (1986); I. Hwang et al., *Plant J.* 1, 367 (1991); P. Jarvis et al., *Plant Mol. Biol.* 24, 685 (1994); L. Le Guen et al., *Mol. Gen. Genet.* 245, 390 (1994).
3. D. Shalon, thesis, Stanford University (1995); and P. O. Brown, in preparation. Microarrays were fabricated on poly-L-lysine-coated microscope slides (Sigma) with a custom-built arraying machine fitted with one printing tip. The tip loaded 1 μ l of PCR product (0.5 mg/ml) from 96-well microtiter plates and deposited ~0.005 μ l per slide on 40 slides at a spacing of 500 μ m. The printed slides were rehydrated for 2 hours in a humid chamber, snap-dried at 100°C for 1 min, rinsed in 0.1% SDS, and treated with 0.05% succinic anhydride prepared in buffer consisting of 50% 1-methyl-2-pyrrolidone and 50% boric acid. The cDNA on the slides was denatured in distilled water for 2 min at 90°C immediately before use. Microarrays were scanned with a laser fluorescent scanner that contained a computer-controlled XY stage and a microscope objective. A mixed gas, multiline laser allowed sequential excitation of the two fluorophores. Emitted light was split according to wavelength and detected with two photomultiplier tubes. Signals were read into a PC with the use of a 12-bit analog-to-digital board. Additional details of microarray fabrication and use may be obtained by means of e-mail (pbrown@crgm.stanford.edu).
4. F. M. Ausubel et al., Eds., *Current Protocols in Molecular Biology* (Greene & Wiley Interscience, New York, 1994), pp. 4.3.1–4.3.4.
5. Polyadenylated (poly(A)⁺) mRNA was prepared from total RNA with the use of Oligotex-dT resin (Qiagen). Reverse transcription (RT) reactions were carried out with a StrataScript RT-PCR kit (Stratagene) modified as follows: 50- μ l reactions contained 0.1 μ g/ μ l of Arabidopsis mRNA, 0.1 ng/ μ l of human AChR mRNA, 0.05 μ g/ μ l of oligo(dT) (21-mer), 1 \times first strand buffer, 0.03 U/ μ l of ribonuclease block, 500 μ M deoxyadenosine triphosphate (dATP), 500 μ M deoxyguanosine triphosphate, 500 μ M dTTP, 40 μ M deoxycytosine triphosphate (dCTP), 40 μ M fluorescein-12-dCTP (or lissamine-5-dCTP), and 0.03 U/ μ l of StrataScript reverse transcriptase. Reactions were incubated for 60 min at 37°C, precipitated with ethanol, and resuspended in 10 μ l of TE (10 mM Tris-HCl and 1 mM EDTA, pH 8.0). Samples were then heated for 3 min at 94°C and chilled on ice. The RNA was degraded by adding 0.25 μ l of 10 N NaOH followed by a 10-min incubation at 37°C. The samples were neutralized by addition of 2.5 μ l of 1 M Tris-Cl (pH 8.0) and 0.25 μ l of 10 N HCl and precipitated with ethanol. Pellets were washed with 70% ethanol, dried to completion in a speedvac, resuspended in 10 μ l of H₂O, and reduced to 3.0 μ l in a speedvac. Fluorescent nucleotide analogs were obtained from New England Nuclear (DuPont).
6. Hybridization reactions contained 1.0 μ l of fluorescent cDNA synthesis product (5) and 1.0 μ l of hybridization buffer (10 \times saline sodium citrate (SSC) and 0.2% SDS). The 2.0- μ l probe mixtures were aliquoted onto the microarray surface and covered with cover slips (12 mm round). Arrays were transferred to a hybridization chamber (3) and incubated for 18 hours at 65°C. Arrays were washed for 5 min at room temperature (25°C) in low-stringency wash buffer (1 \times SSC and 0.1% SDS), then for 10 min at room temperature in high-stringency wash buffer (0.1 \times SSC and 0.1% SDS). Arrays were scanned in 0.1 \times SSC with the use of a fluorescence laser-scanning device (7).
7. Samples of poly(A)⁺ mRNA (4, 5) were spotted onto nylon membranes (Nytran) and crosslinked with ultraviolet light with the use of a Stratalinker 1800 (Stratagene). Probes were prepared by random priming with the use of a Prime-It II kit (Stratagene) in the presence of [³²P]dATP. Hybridizations were carried out according to the instructions of the manufacturer. Quantitation was performed on a PhosphorImager (Molecular Dynamics).
8. M. Schena and R. W. Davis, *Proc. Natl. Acad. Sci. U.S.A.* 89, 3894 (1992); M. Schena, A. M. Lloyd, R. W. Davis, *Genes Dev.* 7, 367 (1993); M. Schena and R. W. Davis, *Proc. Natl. Acad. Sci. U.S.A.* 91, 8393 (1994).
9. H. Hofte et al., *Plant J.* 4, 1051 (1993); T. Newman et al., *Plant Physiol.* 106, 1241 (1994).
10. N. E. Morton, *Proc. Natl. Acad. Sci. U.S.A.* 88, 7474 (1991); E. D. Green and R. H. Waterston, *J. Am. Med. Assoc.* 266, 1966 (1991); C. Belletre-Chantelot, *Cell* 70, 1059 (1992); D. R. Cox et al., *Science* 265, 2031 (1994).
11. E. S. Kawasaki et al., *Proc. Natl. Acad. Sci. U.S.A.* 85, 5698 (1988).
12. The laser fluorescent scanner was designed and fabricated in collaboration with S. Smith of Stanford University. Scanner and analysis software was developed by R. X. Xia. The succinic anhydride reaction was suggested by J. Mulligan and J. Van Ness of Darwin Molecular Corporation. Thanks to S. Theologis, C. Somerville, K. Yamamoto, and members of the laboratories of R.W.D. and P.O.B. for critical comments. Supported by the Howard Hughes Medical Institute and by grants from NIH (R21HG00450) (P.O.B.) and R37AG00198 (R.W.D.) and from NSF (MCB9106011) (R.W.D.) and by an NSF graduate fellowship (D.S.). P.O.B. is an assistant investigator of the Howard Hughes Medical Institute.

11 August 1995; accepted 22 September 1995

Gene Therapy in Peripheral Blood Lymphocytes and Bone Marrow for ADA⁻ Immunodeficient Patients

Claudio Bordignon,* Luigi D. Notarangelo, Nadia Nobili, Giuliana Ferrari, Giulia Casorati, Paola Panina, Evelina Mazzolari, Daniela Maggioni, Claudia Rossi, Paolo Servida, Alberto G. Ugazio, Fulvio Mavilio

Adenosine deaminase (ADA) deficiency results in severe combined immunodeficiency, the first genetic disorder treated by gene therapy. Two different retroviral vectors were used to transfer ex vivo the human ADA minigene into bone marrow cells and peripheral blood lymphocytes from two patients undergoing exogenous enzyme replacement therapy. After 2 years of treatment, long-term survival of T and B lymphocytes, marrow cells, and granulocytes expressing the transferred ADA gene was demonstrated and resulted in normalization of the immune repertoire and restoration of cellular and humoral immunity. After discontinuation of treatment, T lymphocytes, derived from transduced peripheral blood lymphocytes, were progressively replaced by marrow-derived T cells in both patients. These results indicate successful gene transfer into long-lasting progenitor cells, producing a functional multilineage progeny.

Severe combined immunodeficiency associated with inherited deficiency of ADA (1) is usually fatal unless affected children are kept in protective isolation or the immune system is reconstituted by bone marrow transplantation from a human leukocyte antigen (HLA)-identical sibling donor (2). This is the therapy of choice, although it is available only for a minority of patients. In recent years, other forms of therapy have been developed, including transplants from haploidentical donors (3, 4), exogenous enzyme replacement (5), and somatic-cell gene therapy (6–9).

We previously reported a preclinical model in which ADA gene transfer and expression

successfully restored immune functions in human ADA-deficient (ADA⁻) peripheral blood lymphocytes (PBLs) in immunodeficient mice in vivo (10, 11). On the basis of these preclinical results, the clinical application of gene therapy for the treatment of ADA⁻ SCID (severe combined immunodeficiency disease) patients who previously failed exogenous enzyme replacement therapy was approved by our Institutional Ethical Committees and by the Italian National Committee for Bioethics (12). In addition to evaluating the safety and efficacy of the gene therapy procedure, the aim of the study was to define the relative role of PBLs and hematopoietic stem cells in the long-term reconstitution of immune functions after retroviral vector-mediated ADA gene transfer. For this purpose, two structurally identical vectors expressing the human ADA complementary DNA (cDNA), distinguishable by the presence of alternative restriction sites in a nonfunctional region of the viral long-terminal repeat (LTR), were used to transduce PBLs and bone marrow (BM) cells independently. This procedure allowed identification of the origin of

C. Bordignon, N. Nobili, G. Ferrari, D. Maggioni, C. Rossi, P. Servida, F. Mavilio, Telethon Gene Therapy Program for Genetic Diseases, DIBIT, Istituto Scientifico H. S. Raffaele, Milan, Italy.

L. D. Notarangelo, E. Mazzolari, A. G. Ugazio, Department of Pediatrics, University of Brescia Medical School, Brescia, Italy.

G. Casorati, Unità di Immunochimica, DIBIT, Istituto Scientifico H. S. Raffaele, Milan, Italy.

P. Panina, Roche Milano Ricerche, Milan, Italy.

*To whom correspondence should be addressed.

REPORTS

Axl1p sequence following Ser²⁰⁰ and occurs within the domain of Axl1p that shows homology with hDE (14). To delete the complete STE23 sequence and create the *ste23Δ::URA3* mutation, polymerase chain reaction (PCR) primers (5'-TCGGAAGACCTCAT-TCTTGCTCATTTGATATTGCTC-TGTAGATTG-TACTGAGAGTGCAC-3'; and 5'-GCTACAAACAGC-GTGGACTTGAATGCCCGACATCTTCGACTGT-GCGGTATTTCACACCG-3') were used to amplify the URA3 sequence of pRS316, and the reaction product was transformed into yeast for one-step gene replacement [R. Rothstein, *Methods Enzymol.* 194, 281 (1991)]. To create the *axl1Δ::LEU2* mutation contained on p114, a 5.0-kb *SacI* fragment from pAXL1 was cloned into pUC19, and an internal 4.0-kb *HpaI*-*XhoI* fragment was replaced with a *LEU2* fragment. To construct the *ste23Δ::LEU2* allele (a deletion corresponding to 931 amino acids) carried on p153, a *LEU2* fragment was used to replace the 2.8-kb *PmlI*-*Ecd136II* fragment of STE23, which occurs within a 6.2-kb *HindIII*-*BglII* genomic fragment carried on pSP72 (Promega). To create YEpMFA1, a 1.6-kb *BamHI* fragment containing MFA1, from pKX16 [K. Kuchler, R. E. Sterne, J. Thormer, *EMBO J.* 8, 3973 (1989)], was ligated into the *BamHI* site of YEp351 [J. E. Hill, A. M. Myers, T. J. Koerner, A. Tzagoloff, *Yeast* 2, 163 (1986)].

uct. pC225 is a KS+ (Stratagene) plasmid containing a 0.5-kb *BamHI*-*SstI* fragment from pAXL1. Substitution mutations of the proposed active site of Axl1p were created with the use of pC225 and site-specific mutagenesis involving appropriate synthetic oligonucleotides [*axl1-H68A*, 5'-GTGCTCACAAAGCGCT-GCCAAACCGGC-3'; *axl1-E71A*, 5'-AAGAATCAT-GTGCGCACAAAGGTGCGC-3'; and *axl1-E71D*, 5'-AAGAATCATGTGATCACAAAGGTGCGC-3']. The mutations were confirmed by sequence analysis. After mutagenesis, the 0.4-kb *BamHI*-*MscI* fragment from the mutagenized pC225 plasmids was transferred into pAXL1 to create a set of pRS316 plasmids carrying different AXL1 alleles, p124 (*axl1-H68A*), p130 (*axl1-E71A*), and p132 (*axl1-E71D*). Similarly, a set of HA-tagged alleles carried on YEp352 were created after replacement of the p151 *BamHI*-*MscI* fragment, to generate p161 (*axl1-E71A*), p162 (*axl1-*

32

N. Davis, T. Favero, C. de Hoog, and S. Kim for comments on the manuscript. Supported by a grant to C.B. from the Natural Sciences and Engineering Research Council of Canada. Support for M.N.A. was from a California Tobacco-Related Disease Research Program postdoctoral fellowship (4FT-0083).

22 June 1995; accepted 21 August 1995

Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray

Mark Schena,* Dari Shalon,*† Ronald W. Davis, Patrick O. Brown†

A high-capacity system was developed to monitor the expression of many genes in parallel. Microarrays prepared by high-speed robotic printing of complementary DNAs on glass were used for quantitative expression measurements of the corresponding genes. Because of the small format and high density of the arrays, hybridization volumes of 2 microliters could be used that enabled detection of rare transcripts in probe mixtures derived from 2 micrograms of total cellular messenger RNA. Differential expression measurements of 45 *Arabidopsis* genes were made by means of simultaneous, two-color fluorescence hybridization.

The temporal, developmental, topographical, histological, and physiological patterns in which a gene is expressed provide clues to its biological role. The large and expanding database of complementary DNA (cDNA) sequences from many organisms (1) presents the opportunity of defining these patterns at the level of the whole genome.

For these studies, we used the small flowering plant *Arabidopsis thaliana* as a model organism. *Arabidopsis* possesses many advantages for gene expression analysis, including the fact that it has the smallest genome of any higher eukaryote examined to date (2). Forty-five cloned *Arabidopsis* cDNAs (Table 1), including 14 complete sequences and 31 expressed sequence tags (ESTs), were used as gene-specific targets. We obtained the ESTs by selecting cDNA clones at random from an *Arabidopsis* cDNA library. Sequence analysis revealed that 28 of the 31 ESTs matched sequences

in the database (Table 1). Three additional cDNAs from other organisms served as controls in the experiments.

The 48 cDNAs, averaging ~1.0 kb, were amplified with the polymerase chain reaction (PCR) and deposited into individual wells of a 96-well microtiter plate. Each sample was duplicated in two adjacent wells to allow the reproducibility of the arraying and hybridization process to be tested. Samples from the microtiter plate were printed onto glass microscope slides in an area measuring 3.5 mm by 5.5 mm with the use of a high-speed arraying machine (3). The arrays were processed by chemical and heat treatment to attach the DNA sequences to the glass surface and denature them (3). Three arrays, printed in a single lot, were used for the experiments here. A single microtiter plate of PCR products provides sufficient material to print at least 500 arrays.

Fluorescent probes were prepared from total *Arabidopsis* mRNA (4) by a single round of reverse transcription (5). The *Arabidopsis* mRNA was supplemented with human acetylcholine receptor (AChR) mRNA at a dilution of 1:10,000 (w/w) before cDNA synthesis, to provide an internal standard for calibration (5). The resulting fluorescently labeled cDNA mixture was hybridized to an array at high stringency (6) and scanned

24. J. Chant and I. Herskowitz, *Cell* 65, 1203 (1991).
25. B. W. Matthews, *Acc. Chem. Res.* 21, 333 (1988).
26. K. Kuchler, H. G. Dohlman, J. Thormer, *J. Cell Biol.* 120, 1203 (1993); R. Kolling and C. P. Hollenberg, *EMBO J.* 13, 3261 (1994); C. Berkower, D. Loayza, S. Michaelis, *Mol. Biol. Cell* 5, 1185 (1994).
27. A. Bender and J. R. Pringle, *Proc. Natl. Acad. Sci. U.S.A.* 86, 9976 (1989); J. Chant, K. Corrado, J. R. Pringle, I. Herskowitz, *Cell* 65, 1213 (1991); S. Powers, E. Gonzales, T. Christensen, J. Cubert, D. Broek, *ibid.*, p. 1225; H. O. Park, J. Chant, I. Herskowitz, *Nature* 365, 269 (1993); J. Chant, *Trends Genet.* 10, 328 (1994); _____ and J. R. Pringle, *J. Cell Biol.* 129, 751 (1995); J. Chant, M. Mischke, E. Mitchell, I. Herskowitz, J. R. Pringle, *ibid.*, p. 767.
28. G. F. Sprague Jr., *Methods. Enzymol.* 194, 77 (1991).
29. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
30. A W303 1A derivative, SY2625 (MATa *ura3-1 leu2-3, 112 trp1-1 ade2-1 can1-100 ss1Δ mfa2Δ::FUS1-lacZ his3Δ::FUS1-HIS3*), was the parent strain for the mutant search. SY2625 derivatives for the mating assays, secreted pheromone assays, and the pulse-chase experiments included the following strains: Y49 (*ste22-1*), Y115 (*mfa1Δ::LEU2*), Y142 (*axl1::URA3*), Y173 (*axl1Δ::LEU2*), Y220 (*axl1::URA3 ste23Δ::URA3*), Y221 (*ste23Δ::URA3*), Y231 (*axl1Δ::LEU2 ste23Δ::LEU2*), and Y233 (*ste23Δ::LEU2*). MATa derivatives of SY2625 included the following strains: Y199 (SY2625 made MATa), Y278 (*ste22-1*), Y195 (*mfa1Δ::LEU2*), Y196 (*axl1Δ::LEU2*), and Y197 (*axl1::URA3*). The EG123 (MATa *leu2 ura3 trp1 can1 his4*) genetic background was used to create a set of strains for analysis of bud site selection. EG123 derivatives included the following strains: Y175 (*axl1Δ::LEU2*), Y223 (*axl1::URA3*), Y234 (*ste23Δ::LEU2*), and Y272 (*axl1Δ::LEU2 ste23Δ::LEU2*). MATa derivatives of EG123 included the following strains: Y214 (EG123 made MATa) and Y293 (*axl1Δ::LEU2*). All strains were generated by means of standard genetic or molecular methods involving the appropriate constructs (23). In particular, the *axl1 ste23* double mutant strains were created by crossing of the appropriate MATa *ste23* and MATa *axl1* mutants, followed by sporulation of the resultant diploid and isolation of the double mutant from nonparental di-type tetrads. Gene disruptions were confirmed with either PCR or Southern (DNA) analysis.
31. p129 is a YEp352 [J. E. Hill, A. M. Myers, T. J. Koerner, A. Tzagoloff, *Yeast* 2, 163 (1986)] plasmid containing a 5.5-kb *SacI* fragment of pAXL1. p151 was derived from p129 by insertion of a linker at the *BglII* site within AXL1, which led to an in-frame insertion of the hemagglutinin (HA) epitope (DQTPYDVPDYA) (29) between amino acids 854 and 855 of the AXL1 prod-

M. Schena and R. W. Davis, Department of Biochemistry, Beckman Center, Stanford University Medical Center, Stanford, CA 94305, USA.
D. Shalon and P. O. Brown, Department of Biochemistry and Howard Hughes Medical Institute, Beckman Center, Stanford University Medical Center, Stanford, CA 94305, USA.

*These authors contributed equally to this work.
†Present address: Syntex, Palo Alto, CA 94303, USA.
‡To whom correspondence should be addressed. E-mail: pbrown@cmgm.stanford.edu

with a laser (3). A high-sensitivity scan gave signals that saturated the detector at nearly all of the *Arabidopsis* target sites (Fig. 1A). Calibration relative to the AChR mRNA standard (Fig. 1A) established a sensitivity limit of $\sim 1:50,000$. No detectable hybridization was observed to either the rat glucocorticoid receptor (Fig. 1A) or the yeast TRP4 (Fig. 1A) targets even at the highest scanning sensitivity. A moderate-sensitivity scan

of the same array allowed linear detection of the more abundant transcripts (Fig. 1B). Quantitation of both scans revealed a range of expression levels spanning three orders of magnitude for the 45 genes tested (Table 2). RNA blots (7) for several genes (Fig. 2) corroborated the expression levels measured with the microarray to within a factor of 5 (Table 2).

Differential gene expression was investi-

gated with a simultaneous, two-color hybridization scheme, which served to minimize experimental variation inherent in the comparison of independent hybridizations. Fluorescent probes were prepared from two mRNA sources with the use of reverse transcriptase in the presence of fluorescein- and lissamine-labeled nucleotide analogs, respectively (5). The two probes were then mixed together in equal proportions, hybridized to a single array, and scanned separately for fluorescein and lissamine emission after independent excitation of the two fluorophores (3).

To test whether overexpression of a single gene could be detected in a pool of total *Arabidopsis* mRNA, we used a microarray to analyze a transgenic line overexpressing the single transcription factor HAT4 (8). Fluorescent probes representing mRNA from wild-type and HAT4-transgenic plants were labeled with fluorescein and lissamine, respectively; the two probes were then mixed and hybridized to a single array. An intense hybridization signal was observed at the position of the HAT4 cDNA in the lissamine-specific scan (Fig. 1D), but not in the fluorescein-specific scan of the same array (Fig. 1C). Calibration with AChR mRNA added to the fluorescein and lissamine cDNA synthesis reactions at dilutions of 1:10,000 (Fig. 1C) and 1:100 (Fig. 1D), respectively, revealed a 50-fold elevation of HAT4 mRNA in the transgenic line relative to its abundance in wild-type plants (Table 2). This magnitude of HAT4 overexpression matched that inferred from the Northern (RNA) analysis within a factor of 2 (Fig. 2 and Table 2). Expression of all the other genes monitored on the array differed by less than a factor of 5 between HAT4-transgenic and wild-type plants (Fig. 1, C

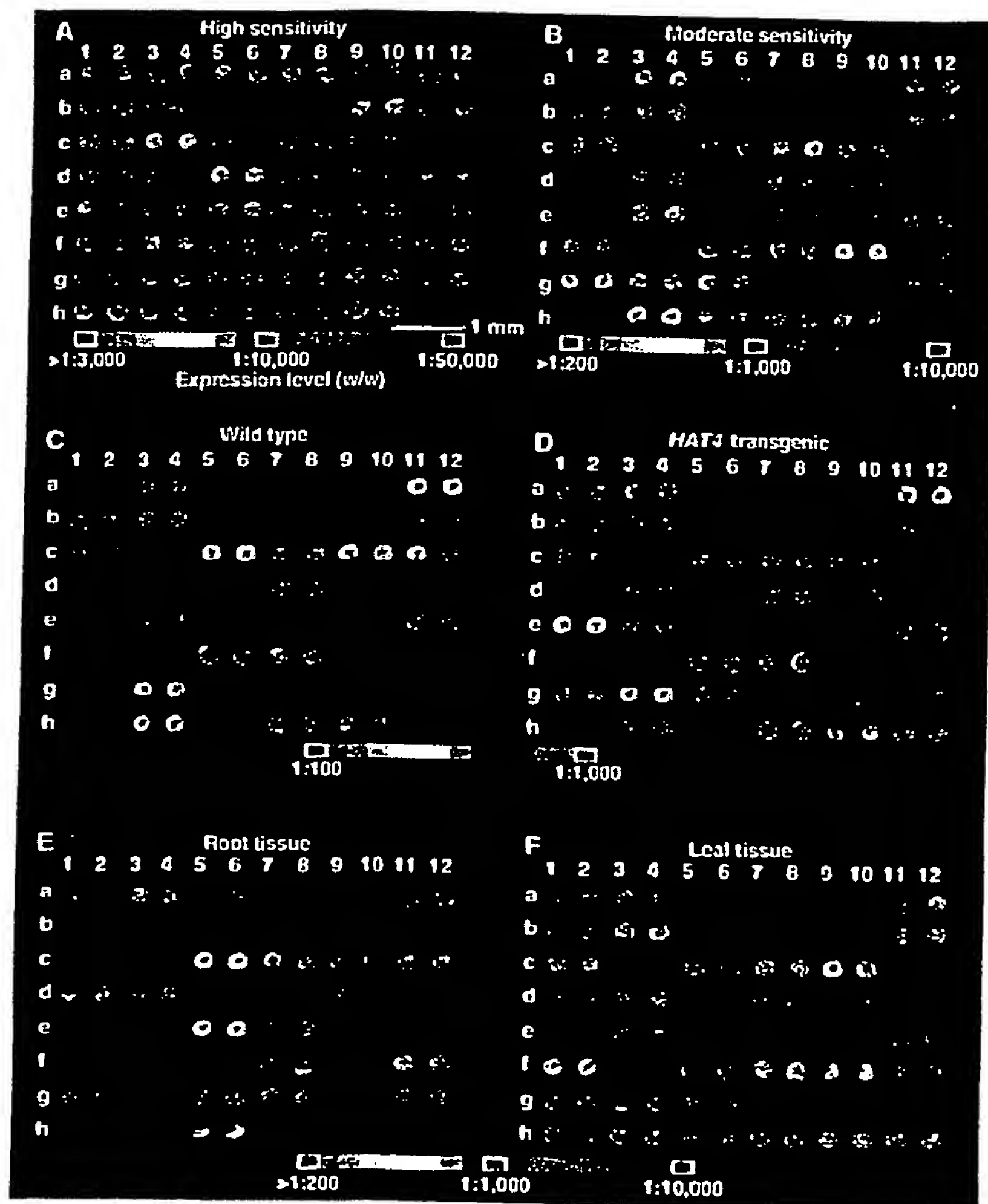


Fig. 1. Gene expression monitored with the use of cDNA microarrays. Fluorescent scans represented in pseudocolor correspond to hybridization intensities. Color bars were calibrated from the signal obtained with the use of known concentrations of human AChR mRNA in independent experiments. Numbers and letters on the axes mark the position of each cDNA. (A) High-sensitivity fluorescein scan after hybridization with fluorescein-labeled cDNA derived from wild-type plants. (B) Same array as in (A) but scanned at moderate sensitivity. (C and D) A single array was probed with a 1:1 mixture of fluorescein-labeled cDNA from wild-type plants and lissamine-labeled cDNA from HAT4-transgenic plants. The single array was then scanned successively to detect the fluorescein fluorescence corresponding to mRNA from wild-type plants (C) and the lissamine fluorescence corresponding to mRNA from HAT4-transgenic plants (D). (E and F) A single array was probed with a 1:1 mixture of fluorescein-labeled cDNA from root tissue and lissamine-labeled cDNA from leaf tissue. The single array was then scanned successively to detect the fluorescein fluorescence corresponding to mRNAs expressed in roots (E) and the lissamine fluorescence corresponding to mRNAs expressed in leaves (F).

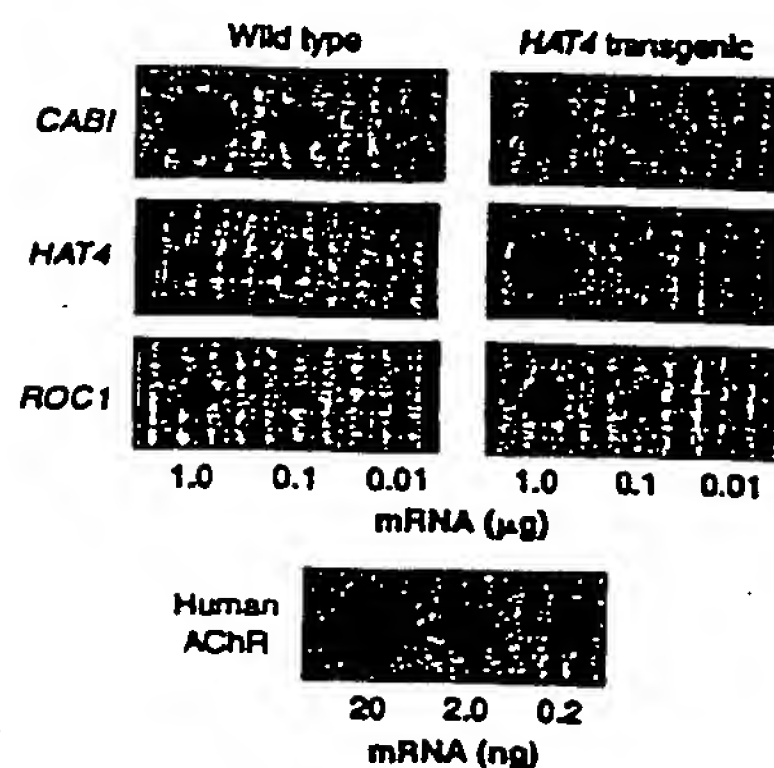


Fig. 2. Gene expression monitored with RNA (Northern) blot analysis. Designated amounts of mRNA from wild-type and HAT4-transgenic plants were spotted onto nylon membranes and probed with the cDNAs indicated. Purified human AChR mRNA was used for calibration.

and D, and Table 2). Hybridization of fluorescein-labeled glucocorticoid receptor cDNA (Fig. 1C) and lissamine-labeled TRP4 cDNA (Fig. 1D) verified the presence of the negative control targets and the lack of optical cross talk between the two fluorophores.

To explore a more complex alteration in expression patterns, we performed a second two-color hybridization experiment with fluorescein- and lissamine-labeled probes prepared from root and leaf mRNA, respectively. The scanning sensitivities for the two fluorophores were normalized by matching the signals resulting from AChR

mRNA, which was added to both cDNA synthesis reactions at a dilution of 1:1000 (Fig. 1, E and F). A comparison of the scans revealed widespread differences in gene expression between root and leaf tissue (Fig. 1, E and F). The mRNA from the light-regulated *CAB1* gene was ~500-fold more abundant in leaf (Fig. 1F) than in root tissue (Fig. 1E). The expression of 26 other genes differed between root and leaf tissue by more than a factor of 5 (Fig. 1, E and F).

The *HAT4*-transgenic line we examined has elongated hypocotyls, early flowering, poor germination, and altered pigmentation (8). Although changes in expression were

observed for *HAT4*, large changes in expression were not observed for any of the other 44 genes we examined. This was somewhat surprising, particularly because comparative analysis of leaf and root tissue identified 27 differentially expressed genes. Analysis of an expanded set of genes may be required to identify genes whose expression changes upon *HAT4* overexpression; alternatively, a comparison of mRNA populations from specific tissues of wild-type and *HAT4*-transgenic plants may allow identification of downstream genes.

At the current density of robotic printing, it is feasible to scale up the fabrication process to produce arrays containing 20,000 cDNA targets. At this density, a single array would be sufficient to provide gene-specific targets encompassing nearly the entire repertoire of expressed genes in the *Arabidopsis* genome (2). The availability of 20,274 ESTs from *Arabidopsis* (1, 9) would provide a rich source of templates for such studies.

The estimated 100,000 genes in the human genome (10) exceeds the number of *Arabidopsis* genes by a factor of 5 (2). This modest increase in complexity suggests that similar cDNA microarrays, prepared from the rapidly growing repertoire of human ESTs (1), could be used to determine the expression patterns of tens of thousands of human genes in diverse cell types. Coupling an amplification strategy to the reverse transcription reaction (11) could make it feasible to monitor expression even in minute tissue samples. A wide variety of acute and chronic physiological and pathological conditions might lead to characteristic changes in the patterns of gene expression in peripheral blood cells or other easily sampled tissues. In concert with cDNA microarrays for monitoring complex expression patterns, these tissues might therefore serve as sensitive *in vivo* sensors for clinical diagnosis. Microarrays of cDNAs could thus provide a useful link between human gene sequences and clinical medicine.

Table 2. Gene expression monitoring by microarray and RNA blot analyses; tg, *HAT4*-transgenic. See Table 1 for additional gene information. Expression levels (w/w) were calibrated with the use of known amounts of human AChR mRNA. Values for the microarray were determined from microarray scans (Fig. 1); values for the RNA blot were determined from RNA blots (Fig. 2).

Gene	Expression level (w/w)	
	Microarray	RNA blot
<i>CAB1</i>	1:48	1:83
<i>CAB1</i> (tg)	1:120	1:150
<i>HAT4</i>	1:8300	1:6300
<i>HAT4</i> (tg)	1:150	1:210
<i>ROC1</i>	1:1200	1:1800
<i>ROC1</i> (tg)	1:260	1:1300

Table 1. Sequences contained on the cDNA microarray. Shown is the position, the known or putative function, and the accession number of each cDNA in the microarray (Fig. 1). All but three of the ESTs used in this study matched a sequence in the database. NADH, reduced form of nicotinamide adenine dinucleotide; ATPase, adenosine triphosphatase; GTP, guanosine triphosphate.

Position	cDNA	Function	Accession number
a1, 2	AChR	Human AChR	.
a3, 4	EST3	Actin	H36236
a5, 6	EST6	NADH dehydrogenase	Z27010
a7, 8	AAC1	Actin 1	M20016
a9, 10	EST12	Unknown	U36594†
a11, 12	EST13	Actin	T45783
b1, 2	<i>CAB1</i>	Chlorophyll a/b binding	M85150
b3, 4	EST17	Phosphoglycerate kinase	T44490
b5, 6	GA4	Gibberellic acid biosynthesis	L37126
b7, 8	EST19	Unknown	U36595†
b9, 10	<i>GBF-1</i>	G-box binding factor 1	X63894
b11, 12	EST23	Elongation factor	X52256
c1, 2	EST29	Aldolase	T04477
c3, 4	<i>GBF-2</i>	G-box binding factor 2	X63895
c5, 6	EST34	Chloroplast protease	R87034
c7, 8	EST35	Unknown	T14152
c9, 10	EST41	Catalase	T22720
c11, 12	rGR	Rat glucocorticoid receptor	M14053
d1, 2	EST42	Unknown	U36596†
d3, 4	EST45	ATPase	J04185
d5, 6	<i>HAT1</i>	Homeobox-leucine zipper 1	U09332
d7, 8	EST46	Light harvesting complex	T04063
d9, 10	EST49	Unknown	T76267
d11, 12	<i>HAT2</i>	Homeobox-leucine zipper 2	U09335
e1, 2	<i>HAT4</i>	Homeobox-leucine zipper 4	M90394
e3, 4	EST50	Phosphoribulokinase	T04344
e5, 6	<i>HAT5</i>	Homeobox-leucine zipper 5	M90416
e7, 8	EST51	Unknown	Z33675
e9, 10	<i>HAT22</i>	Homeobox-leucine zipper 22	U09336
e11, 12	EST52	Oxygen evolving	T21749
f1, 2	EST59	Unknown	Z34607
f3, 4	<i>KNAT1</i>	Knotted-like homeobox 1	U14174
f5, 6	EST60	RuBisCO small subunit	X14564
f7, 8	EST69	Translation elongation factor	T42799
f9, 10	<i>PPH1</i>	Protein phosphatase 1	U34803
f11, 12	EST70	Unknown	T44621
g1, 2	EST75	Chloroplast protease	T43698
g3, 4	EST78	Unknown	R65481
g5, 6	<i>ROC1</i>	Cyclophilin	L14844
g7, 8	EST82	GTP binding	X59152
g9, 10	EST83	Unknown	Z33795
g11, 12	EST84	Unknown	T45278
h1, 2	EST91	Unknown	T13832
h3, 4	EST96	Unknown	R64816
h5, 6	<i>SAR1</i>	Synaptobrevin	M90418
h7, 8	EST100	Light harvesting complex	Z18205
h9, 10	EST103	Light harvesting complex	X03909
h11, 12	<i>TRP4</i>	Yeast tryptophan biosynthesis	X04273

*Proprietary sequence of Stratagene (La Jolla, California).

†No match in the database; novel EST.

REFERENCES AND NOTES

1. The current EST database (dbEST release 091495) from the National Center for Biotechnology Information (Bethesda, MD) contains a total of 322,225 entries, including 255,845 from the human genome and 21,044 from Arabidopsis. Access is available via the World Wide Web (<http://www.ncbi.nlm.nih.gov>).
2. E. M. Meyerowitz and R. E. Pruitt, *Science* 229, 1214 (1985); R. E. Pruitt and E. M. Meyerowitz, *J. Mol. Biol.* 187, 169 (1986); L. Hwang et al., *Plant J.* 1, 367 (1991); P. Jarvis et al., *Plant Mol. Biol.* 24, 685 (1994); L. Le Guen et al., *Mol. Gen. Genet.* 245, 390 (1994).
3. D. Sharon, thesis, Stanford University (1995); and P. O. Brown, in preparation. Microarrays were fabricated on poly-L-lysine-coated microscope slides (Sigma) with a custom-built arraying machine fitted with one printing tip. The tip loaded 1 μ l of PCR product (0.5 mg/ml) from 96-well microtiter plates and deposited \sim 0.005 μ l per slide on 40 slides at a spacing of 500 μ m. The printed slides were rehydrated for 2 hours in a humid chamber, snap-dried at 100°C for 1 min, rinsed in 0.1% SDS, and treated with 0.05% succinic anhydride prepared in buffer consisting of 50% 1-methyl-2-pyrrolidinone and 50% boric acid. The cDNA on the slides was denatured in distilled water for 2 min at 90°C immediately before use. Microarrays were scanned with a laser fluorescent scanner that contained a computer-controlled XY stage and a microscope objective. A mixed gas, multiline laser allowed sequential excitation of the two fluorophores. Emitted light was split according to wavelength and detected with two photomultiplier tubes. Signals were read into a PC with the use of a 12-bit analog-to-digital board. Additional details of microarray fabrication and use may be obtained by means of e-mail (pbrown@crgm.stanford.edu).
4. F. M. Ausubel et al., Eds., *Current Protocols in Molecular Biology* (Greene & Wiley Interscience, New York, 1994), pp. 4.3.1–4.3.4.
5. Polyadenylated [poly(A)⁺] mRNA was prepared from total RNA with the use of Oligotex-dT resin (Qiagen). Reverse transcription (RT) reactions were carried out with a StrataScript RT-PCR kit (Stratagene) modified as follows: 50- μ l reactions contained 0.1 μ g/ μ l of Arabidopsis mRNA, 0.1 ng/ μ l of human AChR mRNA, 0.05 μ g/ μ l of oligo(dT) (21-mer), 1 \times first strand buffer, 0.03 U/ μ l of ribonuclease block, 500 μ M deoxyadenosine triphosphate (dATP), 500 μ M deoxyguanosine triphosphate, 500 μ M dTTP, 40 μ M deoxycytosine triphosphate (dCTP), 40 μ M fluorescein-12-dCTP (or lissamine-5-dCTP), and 0.03 U/ μ l of StrataScript reverse transcriptase. Reactions were incubated for 60 min at 37°C, precipitated with ethanol, and resuspended in 10 μ l of TE (10 mM tri-HCl and 1 mM EDTA, pH 8.0). Samples were then heated for 3 min at 94°C and chilled on ice. The RNA was degraded by adding 0.25 μ l of 10 N NaOH followed by a 10-min incubation at 37°C. The samples were neutralized by addition of 2.5 μ l of 1 M Tris-Cl (pH 8.0) and 0.25 μ l of 10 N HCl and precipitated with ethanol. Pellets were washed with 70% ethanol, dried to completion in a speedvac, resuspended in 10 μ l of H₂O, and reduced to 3.0 μ l in a speedvac. Fluorescent nucleotide analogs were obtained from New England Nuclear (DuPont).
6. Hybridization reactions contained 1.0 μ l of fluorescent cDNA synthesis product (5) and 1.0 μ l of hybridization buffer [10 \times saline sodium citrate (SSC) and 0.2% SDS]. The 2.0- μ l probe mixtures were aliquoted onto the microarray surface and covered with cover slips (12 mm round). Arrays were transferred to a hybridization chamber (3) and incubated for 18 hours at 65°C. Arrays were washed for 5 min at room temperature (25°C) in low-stringency wash buffer (1 \times SSC and 0.1% SDS), then for 10 min at room temperature in high-stringency wash buffer (0.1 \times SSC and 0.1% SDS). Arrays were scanned in 0.1 \times SSC with the use of a fluorescence laser-scanning device (3).
7. Samples of poly(A)⁺ mRNA (4, 5) were spotted onto nylon membranes (Nytan) and crosslinked with ultraviolet light with the use of a Stratalinker 1800 (Stratagene). Probes were prepared by random priming with the use of a Prime-It II kit (Stratagene) in the presence of [³²P]dATP. Hybridizations were carried out according to the instructions of the manufacturer. Quantitation was performed on a Phosphorimager (Molecular Dynamics).
8. M. Schena and R. W. Davis, *Proc. Natl. Acad. Sci. U.S.A.* 89, 3894 (1992); M. Schena, A. M. Lloyd, R. W. Davis, *Genes Dev.* 7, 367 (1993); M. Schena and R. W. Davis, *Proc. Natl. Acad. Sci. U.S.A.* 91, 8393 (1994).
9. H. Hofte et al., *Plant J.* 4, 1051 (1993); T. Newman et al., *Plant Physiol.* 106, 1241 (1994).
10. N. E. Morton, *Proc. Natl. Acad. Sci. U.S.A.* 88, 7474 (1991); E. D. Green and R. H. Waterston, *J. Am. Med. Assoc.* 266, 1966 (1991); C. Belletane-Chantelot, *Cell* 70, 1059 (1992); D. R. Cox et al., *Science* 265, 2031 (1994).
11. E. S. Kawasaki et al., *Proc. Natl. Acad. Sci. U.S.A.* 85, 5688 (1988).
12. The laser fluorescent scanner was designed and fabricated in collaboration with S. Smith of Stanford University. Scanner and analysis software was developed by R. X. Xia. The succinic anhydride reaction was suggested by J. Mulligan and J. Van Ness of Darwin Molecular Corporation. Thanks to S. Theologis, C. Somerville, K. Yamamoto, and members of the laboratories of R.W.D. and P.O.B. for critical comments. Supported by the Howard Hughes Medical Institute and by grants from NIH (R21HG00450) (P.O.B.) and R37AG00198 (R.W.D.) and from NSF (MCB9106011) (R.W.D.) and by an NSF graduate fellowship (D.S.). P.O.B. is an assistant investigator of the Howard Hughes Medical Institute.

11 August 1995; accepted 22 September 1995

Gene Therapy in Peripheral Blood Lymphocytes and Bone Marrow for ADA⁻ Immunodeficient Patients

Claudio Bordignon,* Luigi D. Notarangelo, Nadia Nobili, Giuliana Ferrari, Giulia Casorati, Paola Panina, Evelina Mazzolari, Daniela Maggioni, Claudia Rossi, Paolo Servida, Alberto G. Ugazio, Fulvio Mavilio

Adenosine deaminase (ADA) deficiency results in severe combined immunodeficiency, the first genetic disorder treated by gene therapy. Two different retroviral vectors were used to transfer ex vivo the human ADA minigene into bone marrow cells and peripheral blood lymphocytes from two patients undergoing exogenous enzyme replacement therapy. After 2 years of treatment, long-term survival of T and B lymphocytes, marrow cells, and granulocytes expressing the transferred ADA gene was demonstrated and resulted in normalization of the immune repertoire and restoration of cellular and humoral immunity. After discontinuation of treatment, T lymphocytes, derived from transduced peripheral blood lymphocytes, were progressively replaced by marrow-derived T cells in both patients. These results indicate successful gene transfer into long-lasting progenitor cells, producing a functional multilineage progeny.

Severe combined immunodeficiency associated with inherited deficiency of ADA (1) is usually fatal unless affected children are kept in protective isolation or the immune system is reconstituted by bone marrow transplantation from a human leukocyte antigen (HLA)-identical sibling donor (2). This is the therapy of choice, although it is available only for a minority of patients. In recent years, other forms of therapy have been developed, including transplants from haploidentical donors (3, 4), exogenous enzyme replacement (5), and somatic-cell gene therapy (6–9).

We previously reported a preclinical model in which ADA gene transfer and expression

successfully restored immune functions in human ADA-deficient (ADA⁻) peripheral blood lymphocytes (PBLs) in immunodeficient mice in vivo (10, 11). On the basis of these preclinical results, the clinical application of gene therapy for the treatment of ADA⁻ SCID (severe combined immunodeficiency disease) patients who previously failed exogenous enzyme replacement therapy was approved by our Institutional Ethical Committees and by the Italian National Committee for Bioethics (12). In addition to evaluating the safety and efficacy of the gene therapy procedure, the aim of the study was to define the relative role of PBLs and hematopoietic stem cells in the long-term reconstitution of immune functions after retroviral vector-mediated ADA gene transfer. For this purpose, two structurally identical vectors expressing the human ADA complementary DNA (cDNA), distinguishable by the presence of alternative restriction sites in a nonfunctional region of the viral long-terminal repeat (LTR), were used to transduce PBLs and bone marrow (BM) cells independently. This procedure allowed identification of the origin of

C. Bordignon, N. Nobili, G. Ferrari, D. Maggioni, C. Rossi, P. Servida, F. Mavilio, Telethon Gene Therapy Program for Genetic Diseases, DIBIT, Istituto Scientifico H. S. Raffaele, Milan, Italy.

L. D. Notarangelo, E. Mazzolari, A. G. Ugazio, Department of Pediatrics, University of Brescia Medical School, Brescia, Italy.

G. Casorati, Unità di Immunochimica, DIBIT, Istituto Scientifico H. S. Raffaele, Milan, Italy.

P. Panina, Roche Milano Ricerche, Milan, Italy.

*To whom correspondence should be addressed.



INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁶ : G01N 33/543, 33/68	A1	(11) International Publication Number: WO 95/35505 (43) International Publication Date: 28 December 1995 (28.12.95)
(21) International Application Number: PCT/US95/07659 (22) International Filing Date: 16 June 1995 (16.06.95) (30) Priority Data: 08/261,388 17 June 1994 (17.06.94) US 08/477,809 7 June 1995 (07.06.95) US (71) Applicant: THE BOARD OF TRUSTEES OF THE LELAND STANFORD JUNIOR UNIVERSITY [US/US]; Stanford, CA 94305 (US). (72) Inventors: SHALON, Tidhar, Dari; 364 Fletcher Drive, Atherton, CA 94027 (US). BROWN, Patri�k, O.; 76 Peter Coutts Circle, Stanford, CA 94305 (US). (74) Agent: DEHLINGER, Peter, J.; Dehlinger & Associates, P.O. Box 60850, Palo Alto, CA 94306-1546 (US).		(81) Designated States: AU, CA, JP, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE). Published <i>With international search report.</i>
(54) Title: METHOD AND APPARATUS FOR FABRICATING MICROARRAYS OF BIOLOGICAL SAMPLES (57) Abstract A method and apparatus for forming microarrays of biological samples on a support are disclosed. The method involves dispensing a known volume of a reagent at each of a selected array position, by tapping a capillary dispenser on the support under conditions effective to draw a defined volume of liquid onto the support. The apparatus is designed to produce a microarray of such regions in an automated fashion.		

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	GB	United Kingdom	MR	Mauritania
AU	Australia	GE	Georgia	MW	Malawi
BB	Barbados	GN	Guinea	NE	Niger
BE	Belgium	GR	Greece	NL	Netherlands
BF	Burkina Faso	HU	Hungary	NO	Norway
BG	Bulgaria	IE	Ireland	NZ	New Zealand
BJ	Benin	IT	Italy	PL	Poland
BR	Brazil	JP	Japan	PT	Portugal
BY	Belarus	KE	Kenya	RO	Romania
CA	Canada	KG	Kyrgyzstan	RU	Russian Federation
CF	Central African Republic	KP	Democratic People's Republic of Korea	SD	Sudan
CG	Congo	KR	Republic of Korea	SE	Sweden
CH	Switzerland	KZ	Kazakhstan	SI	Slovenia
CI	Côte d'Ivoire	LI	Liechtenstein	SK	Slovakia
CM	Cameroon	LK	Sri Lanka	SN	Senegal
CN	China	LU	Luxembourg	TD	Chad
CS	Czechoslovakia	LV	Latvia	TG	Togo
CZ	Czech Republic	MC	Monaco	TJ	Tajikistan
DE	Germany	MD	Republic of Moldova	TT	Trinidad and Tobago
DK	Denmark	MG	Madagascar	UA	Ukraine
ES	Spain	ML	Mali	US	United States of America
FI	Finland	MN	Mongolia	UZ	Uzbekistan
FR	France			VN	Viet Nam
GA	Gabon				

**METHOD AND APPARATUS FOR FABRICATING
MICROARRAYS OF BIOLOGICAL SAMPLES**

Field of the Invention

5 This invention relates to a method and apparatus for fabricating microarrays of biological samples for large scale screening assays, such as arrays of DNA samples to be used in DNA hybridization assays for genetic research and diagnostic applications.

10

References

Abouzied, et al., *Journal of AOAC International* 77(2):495-500 (1994).

Bohlander, et al., *Genomics* 13:1322-1324 (1992).

15 Drmanac, et al., *Science* 260:1649-1652 (1993).

Fodor, et al., *Science* 251:767-773 (1991).

Khrapko, et al., *DNA Sequence* 1:375-388 (1991).

Kuriyama, et al., AN ISFET BIOSENSOR, APPLIED BIOSENSORS (Donald Wise, Ed.), Butterworths, pp. 93-114 (1989).

20 Lehrach, et al., HYBRIDIZATION FINGERPRINTING IN GENOME MAPPING AND SEQUENCING, GENOME ANALYSIS, VOL 1 (Davies and Tilgham, Eds.), Cold Spring Harbor Press, pp. 39-81 (1990).

Maniatis, et al., MOLECULAR CLONING, A LABORATORY
25 MANUAL, Cold Spring Harbor Press (1989).

Nelson, et al., *Nature Genetics* 4:11-18 (1993).

Pirrung, et al., U.S. Patent No. 5,143,854 (1992).

Riles, et al., *Genetics* 134:81-150 (1993).

Schena, M. et al., *Proc. Nat. Acad. Sci. USA*
89:3894-3898 (1992).

5 Southern, et al., *Genomics* 13:1008-1017 (1992).

Background of the Invention

A variety of methods are currently available for making arrays of biological macromolecules, such as
10 arrays of nucleic acid molecules or proteins. One method for making ordered arrays of DNA on a porous membrane is a "dot blot" approach. In this method, a vacuum manifold transfers a plurality, e.g., 96, aqueous samples of DNA from 3 millimeter diameter wells
15 to a porous membrane. A common variant of this procedure is a "slot-blot" method in which the wells have highly-elongated oval shapes.

The DNA is immobilized on the porous membrane by baking the membrane or exposing it to UV radiation.
20 This is a manual procedure practical for making one array at a time and usually limited to 96 samples per array. "Dot-blot" procedures are therefore inadequate for applications in which many thousand samples must be determined.

25 A more efficient technique employed for making ordered arrays of genomic fragments uses an array of pins dipped into the wells, e.g., the 96 wells of a microtitre plate, for transferring an array of samples to a substrate, such as a porous membrane. One array
30 includes pins that are designed to spot a membrane in a staggered fashion, for creating an array of 9216 spots in a 22 x 22 cm area (Lehrach, et al., 1990). A limitation with this approach is that the volume of DNA spotted in each pixel of each array is highly variable.

In addition, the number of arrays that can be made with each dipping is usually quite small.

An alternate method of creating ordered arrays of nucleic acid sequences is described by Pirrung, et al. (1992), and also by Fodor, et al. (1991). The method involves synthesizing different nucleic acid sequences at different discrete regions of a support. This method employs elaborate synthetic schemes, and is generally limited to relatively short nucleic acid sample, e.g., less than 20 bases. A related method has been described by Southern, et al. (1992).

Khrapko, et al. (1991) describes a method of making an oligonucleotide matrix by spotting DNA onto a thin layer of polyacrylamide. The spotting is done manually with a micropipette.

None of the methods or devices described in the prior art are designed for mass fabrication of microarrays characterized by (i) a large number of micro-sized assay regions separated by a distance of 50-200 microns or less, and (ii) a well-defined amount, typically in the picomole range, of analyte associated with each region of the array.

Furthermore, current technology is directed at performing such assays one at a time to a single array of DNA molecules. For example, the most common method for performing DNA hybridizations to arrays spotted onto porous membrane involves sealing the membrane in a plastic bag (Maniatis, et al., 1989) or a rotating glass cylinder (Robbins Scientific) with the labeled hybridization probe inside the sealed chamber. For arrays made on non-porous surfaces, such as a microscope slide, each array is incubated with the labeled hybridization probe sealed under a coverslip. These techniques require a separate sealed chamber for

each array which makes the screening and handling of many such arrays inconvenient and time intensive.

Abouzied, et al. (1994) describes a method of printing horizontal lines of antibodies on a
5 nitrocellulose membrane and separating regions of the membrane with vertical stripes of a hydrophobic material. Each vertical stripe is then reacted with a different antigen and the reaction between the
10 immobilized antibody and an antigen is detected using a standard ELISA colorimetric technique. Abouzied's technique makes it possible to screen many one-dimensional arrays simultaneously on a single sheet of nitrocellulose. Abouzied makes the nitrocellulose somewhat hydrophobic using a line drawn with PAP Pen
15 (Research Products International). However Abouzied does not describe a technology that is capable of completely sealing the pores of the nitrocellulose. The pores of the nitrocellulose are still physically open and so the assay reagents can leak through the
20 hydrophobic barrier during extended high temperature incubations or in the presence of detergents which makes the Abouzied technique unacceptable for DNA hybridization assays.

Porous membranes with printed patterns of
25 hydrophilic/hydrophobic regions exist for applications such as ordered arrays of bacteria colonies. QA Life Sciences (San Diego CA) makes such a membrane with a grid pattern printed on it. However, this membrane has the same disadvantage as the Abouzied technique since
30 reagents can still flow between the gridded arrays making them unusable for separate DNA hybridization assays.

Pall Corporation make a 96-well plate with a porous filter heat sealed to the bottom of the plate.
35 These plates are capable of containing different

reagents in each well without cross-contamination. However, each well is intended to hold only one target element whereas the invention described here makes a microarray of many biomolecules in each subdivided region of the solid support. Furthermore, the 96 well plates are at least 1 cm thick and prevent the use of the device for many colorimetric, fluorescent and radioactive detection formats which require that the membrane lie flat against the detection surface. The invention described here requires no further processing after the assay step since the barriers elements are shallow and do not interfere with the detection step thereby greatly increasing convenience.

Hyseq Corporation has described a method of making an "array of arrays" on a non-porous solid support for use with their sequencing by hybridization technique. The method described by Hyseq involves modifying the chemistry of the solid support material to form a hydrophobic grid pattern where each subdivided region contains a microarray of biomolecules. Hyseq's flat hydrophobic pattern does not make use of physical blocking as an additional means of preventing cross contamination.

Summary of the Invention

The invention includes, in one aspect, a method of forming a microarray of analyte-assay regions on a solid support, where each region in the array has a known amount of a selected, analyte-specific reagent. The method involves first loading a solution of a selected analyte-specific reagent in a reagent-dispensing device having an elongate capillary channel (i) formed by spaced-apart, coextensive elongate members, (ii) adapted to hold a quantity of the reagent solution and (iii) having a tip region at which aqueous

solution in the channel forms a meniscus. The channel is preferably formed by a pair of spaced-apart tapered elements.

5 The tip of the dispensing device is tapped against a solid support at a defined position on the support surface with an impulse effective to break the meniscus in the capillary channel deposit a selected volume of solution on the surface, preferably a selected volume in the range 0.01 to 100 nl. The two steps are
10 repeated until the desired array is formed.

The method may be practiced in forming a plurality of such arrays, where the solution-depositing step is are applied to a selected position on each of a plurality of solid supports at each repeat cycle.

15 The dispensing device may be loaded with a new solution, by the steps of (i) dipping the capillary channel of the device in a wash solution, (ii) removing wash solution drawn into the capillary channel, and (iii) dipping the capillary channel into the new
20 reagent solution.

Also included in the invention is an automated apparatus for forming a microarray of analyte-assay regions on a plurality of solid supports, where each region in the array has a known amount of a selected,
25 analyte-specific reagent. The apparatus has a holder for holding, at known positions, a plurality of planar supports, and a reagent dispensing device of the type described above.

The apparatus further includes positioning
30 structure for positioning the dispensing device at a selected array position with respect to a support in said holder, and dispensing structure for moving the dispensing device into tapping engagement against a support with a selected impulse effective to deposit a

selected volume on the support, e.g., a selected volume in the volume range 0.01 to 100 nl.

The positioning and dispensing structures are controlled by a control unit in the apparatus. The unit operates to (i) place the dispensing device at a loading station, (ii) move the capillary channel in the device into a selected reagent at the loading station, to load the dispensing device with the reagent, and (iii) dispense the reagent at a defined array position on each of the supports on said holder. The unit may further operate, at the end of a dispensing cycle, to wash the dispensing device by (i) placing the dispensing device at a washing station, (ii) moving the capillary channel in the device into a wash fluid, to load the dispensing device with the fluid, and (iii) remove the wash fluid prior to loading the dispensing device with a fresh selected reagent.

The dispensing device in the apparatus may be one of a plurality of such devices which are carried on the arm for dispensing different analyte assay reagents at selected spaced array positions.

In another aspect, the invention includes a substrate with a surface having a microarray of at least 10^3 distinct polynucleotide or polypeptide biopolymers in a surface area of less than about 1 cm². Each distinct biopolymer (i) is disposed at a separate, defined position in said array, (ii) has a length of at least 50 subunits, and (iii) is present in a defined amount between about 0.1 femtomoles and 100 nanomoles.

In one embodiment, the surface is glass slide surface coated with a polycationic polymer, such as polylysine, and the biopolymers are polynucleotides. In another embodiment, the substrate has a water-impermeable backing, a water-permeable film formed on

the backing, and a grid formed on the film. The grid is composed of intersecting water-impervious grid elements extending from said backing to positions raised above the surface of said film, and partitions the film into a plurality of water-impervious cells. A biopolymer array is formed within each well.

More generally, there is provided a substrate for use in detecting binding of labeled polynucleotides to one or more of a plurality different-sequence, immobilized polynucleotides. The substrate includes, in one aspect, a glass support, a coating of a polycationic polymer, such as polylysine, on said surface of the support, and an array of distinct polynucleotides electrostatically bound non-covalently to said coating, where each distinct biopolymer is disposed at a separate, defined position in a surface array of polynucleotides.

In another aspect, the substrate includes a water-impermeable backing, a water-permeable film formed on the backing, and a grid formed on the film, where the grid is composed of intersecting water-impervious grid elements extending from the backing to positions raised above the surface of the film, forming a plurality of cells. A biopolymer array is formed within each cell.

Also forming part of the invention is a method of detecting differential expression of each of a plurality of genes in a first cell type, with respect to expression of the same genes in a second cell type. In practicing the method, there is first produced fluorescent-labeled cDNA's from mRNA's isolated from the two cells types, where the cDNA'S from the first and second cells are labeled with first and second different fluorescent reporters.

A mixture of the labeled cDNA's from the two cell types is added to an array of polynucleotides

representing a plurality of known genes derived from the two cell types, under conditions that result in hybridization of the cDNA's to complementary-sequence polynucleotides in the array. The array is then
5 examined by fluorescence under fluorescence excitation conditions in which (i) polynucleotides in the array that are hybridized predominantly to cDNA's derived from one of the first and second cell types give a distinct first or second fluorescence emission color,
10 respectively, and (ii) polynucleotides in the array that are hybridized to substantially equal numbers of cDNA's derived from the first and second cell types give a distinct combined fluorescence emission color, respectively. The relative expression of known genes
15 in the two cell types can then be determined by the observed fluorescence emission color of each spot.

These and other objects and features of the invention will become more fully apparent when the following detailed description of the invention is read
20 in conjunction with the accompanying figures.

Brief Description of the Drawings

Fig. 1 is a side view of a reagent-dispensing device having a open-capillary dispensing head
25 constructed for use in one embodiment of the invention;

Figs. 2A-2C illustrate steps in the delivery of a fixed-volume bead on a hydrophobic surface employing the dispensing head from Fig. 1, in accordance with one embodiment of the method of the invention;

30 Fig. 3 shows a portion of a two-dimensional array of analyte-assay regions constructed according to the method of the invention;

Fig. 4 is a planar view showing components of an automated apparatus for forming arrays in accordance
35 with the invention.

Fig. 5 shows a fluorescent image of an actual 20 × 20 array of 400 fluorescently-labeled DNA samples immobilized on a poly-l-lysine coated slide, where the total area covered by the 400 element array is 16 square millimeters;

Fig. 6 is a fluorescent image of a 1.8 cm × 1.8 cm microarray containing lambda clones with yeast inserts, the fluorescent signal arising from the hybridization to the array with approximately half the yeast genome labeled with a green fluorophore and the other half with a red fluorophore;

Fig. 7 shows the translation of the hybridization image of Fig. 6 into a karyotype of the yeast genome, where the elements of Fig.-6 microarray contain yeast DNA sequences that have been previously physically mapped in the yeast genome;

Fig. 8 show a fluorescent image of a 0.5 cm × 0.5 cm microarray of 24 cDNA clones, where the microarray was hybridized simultaneously with total cDNA from wild type *Arabidopsis* plant labeled with a green fluorophore and total cDNA from a transgenic *Arabidopsis* plant labeled with a red fluorophore, and the arrow points to the cDNA clone representing the gene introduced into the transgenic *Arabidopsis* plant;

Fig. 9 shows a plan view of substrate having an array of cells formed by barrier elements in the form of a grid;

Fig. 10 shows an enlarged plan view of one of the cells in the substrate in Fig. 9, showing an array of polynucleotide regions in the cell;

Fig. 11 is an enlarged sectional view of the substrate in Fig. 9, taken along a section line in that figure; and

Fig. 12 is a scanned image of a 3 cm × 3 cm nitrocellulose solid support containing four identical

arrays of M13 clones in each of four quadrants, where each quadrant was hybridized simultaneously to a different oligonucleotide using an open face hybridization method.

5

Detailed Description of the Invention

I. Definitions

Unless indicated otherwise, the terms defined below have the following meanings:

10 "Ligand" refers to one member of a ligand/anti-ligand binding pair. The ligand may be, for example, one of the nucleic acid strands in a complementary, hybridized nucleic acid duplex binding pair; an effector molecule in an effector/receptor binding pair;
15 or an antigen in an antigen/antibody or antigen/antibody fragment binding pair.

"Antiligand" refers to the opposite member of a ligand/anti-ligand binding pair. The antiligand may be the other of the nucleic acid strands in a
20 complementary, hybridized nucleic acid duplex binding pair; the receptor molecule in an effector/receptor binding pair; or an antibody or antibody fragment molecule in antigen/antibody or antigen/antibody fragment binding pair, respectively.

25 "Analyte" or "analyte molecule" refers to a molecule, typically a macromolecule, such as a polynucleotide or polypeptide, whose presence, amount, and/or identity are to be determined. The analyte is one member of a ligand/anti-ligand pair.

30 "Analyte-specific assay reagent" refers to a molecule effective to bind specifically to an analyte molecule. The reagent is the opposite member of a ligand/anti-ligand binding pair.

An "array of regions on a solid support" is a
35 linear or two-dimensional array of preferably discrete

regions, each having a finite area, formed on the surface of a solid support.

A "microarray" is an array of regions having a density of discrete regions of at least about $100/\text{cm}^2$, and preferably at least about $1000/\text{cm}^2$. The regions in a microarray have typical dimensions, e.g., diameters, in the range of between about 10-250 μm , and are separated from other regions in the array by about the same distance.

A support surface is "hydrophobic" if a aqueous-medium droplet applied to the surface does not spread out substantially beyond the area size of the applied droplet. That is, the surface acts to prevent spreading of the droplet applied to the surface by hydrophobic interaction with the droplet.

A "meniscus" means a concave or convex surface that forms on the bottom of a liquid in a channel as a result of the surface tension of the liquid.

"Distinct biopolymers", as applied to the biopolymers forming a microarray, means an array member which is distinct from other array members on the basis of a different biopolymer sequence, and/or different concentrations of the same or distinct biopolymers, and/or different mixtures of distinct or different-concentration biopolymers. Thus an array of "distinct polynucleotides" means an array containing, as its members, (i) distinct polynucleotides, which may have a defined amount in each member, (ii) different, graded concentrations of given-sequence polynucleotides, and/or (iii) different-composition mixtures of two or more distinct polynucleotides.

"Cell type" means a cell from a given source, e.g., a tissue, or organ, or a cell in a given state of

differentiation, or a cell associated with a given pathology or genetic makeup.

II. Method of Microarray Formation

5 This section describes a method of forming a microarray of analyte-assay regions on a solid support or substrate, where each region in the array has a known amount of a selected, analyte-specific reagent.

10 Fig. 1 illustrates, in a partially schematic view, a reagent-dispensing device 10 useful in practicing the method. The device generally includes a reagent dispenser 12 having an elongate open capillary channel 14 adapted to hold a quantity of the reagent solution, such as indicated at 16, as will be described below.

15 The capillary channel is formed by a pair of spaced-apart, coextensive, elongate members 12a, 12b which are tapered toward one another and converge at a tip or tip region 18 at the lower end of the channel. More generally, the open channel is formed by at least two

20 elongate, spaced-apart members adapted to hold a quantity of reagent solutions and having a tip region at which aqueous solution in the channel forms a meniscus, such as the concave meniscus illustrated at 20 in Fig. 2A. The advantages of the open channel

25 construction of the dispenser are discussed below.

 With continued reference to Fig. 1, the dispenser device also includes structure for moving the dispenser rapidly toward and away from a support surface, for effecting deposition of a known amount of solution in

30 the dispenser on a support, as will be described below with reference to Figs. 2A-2C. In the embodiment shown, this structure includes a solenoid 22 which is activatable to draw a solenoid piston 24 rapidly downwardly, then release the piston, e.g., under spring

35 bias, to a normal, raised position, as shown. The

dispenser is carried on the piston by a connecting member 26, as shown. The just-described moving structure is also referred to herein as dispensing means for moving the dispenser into engagement with a solid support, for dispensing a known volume of fluid on the support.

The dispensing device just described is carried on an arm 28 that may be moved either linearly or in an x-y plane to position the dispenser at a selected deposition position, as will be described.

Figs. 2A-2C illustrate the method of depositing a known amount of reagent solution in the just-described dispenser on the surface of a solid support, such as the support indicated at 30. The support is a polymer, glass, or other solid-material support having a surface indicated at 31.

In one general embodiment, the surface is a relatively hydrophilic, *i.e.*, wettable surface, such as a surface having native, bound or covalently attached charged groups. On such surface described below is a glass surface having an absorbed layer of a polycationic polymer, such as poly-l-lysine.

In another embodiment, the surface has or is formed to have a relatively hydrophobic character, *i.e.*, one that causes aqueous medium deposited on the surface to bead. A variety of known hydrophobic polymers, such as polystyrene, polypropylene, or polyethylene have desired hydrophobic properties, as do glass and a variety of lubricant or other hydrophobic films that may be applied to the support surface.

Initially, the dispenser is loaded with a selected analyte-specific reagent solution, such as by dipping the dispenser tip, after washing, into a solution of the reagent, and allowing filling by capillary flow into the dispenser channel. The dispenser is now moved

to a selected position with respect to a support surface, placing the dispenser tip directly above the support-surface position at which the reagent is to be deposited. This movement takes place with the
5 dispenser tip in its raised position, as seen in Fig. 2A, where the tip is typically at least several 1-5 mm above the surface of the substrate.

With the dispenser so positioned, solenoid 22 is now activated to cause the dispenser tip to move
10 rapidly toward and away from the substrate surface, making momentary contact with the surface, in effect, tapping the tip of the dispenser against the support surface. The tapping movement of the tip against the surface acts to break the liquid meniscus in the tip
15 channel, bringing the liquid in the tip into contact with the support surface. This, in turn, produces a flowing of the liquid into the capillary space between the tip and the surface, acting to draw liquid out of the dispenser channel, as seen in Fig. 2B.

20 Fig. 2C shows flow of fluid from the tip onto the support surface, which in this case is a hydrophobic surface. The figure illustrates that liquid continues to flow from the dispenser onto the support surface until it forms a liquid bead 32. At a given bead size,
25 i.e., volume, the tendency of liquid to flow onto the surface will be balanced by the hydrophobic surface interaction of the bead with the support surface, which acts to limit the total bead area on the surface, and by the surface tension of the droplet, which tends
30 toward a given bead curvature. At this point, a given bead volume will have formed, and continued contact of the dispenser tip with the bead, as the dispenser tip is being withdrawn, will have little or no effect on bead volume.

For liquid-dispensing on a more hydrophilic surface, the liquid will have less of a tendency to bead, and the dispensed volume will be more sensitive to the total dwell time of the dispenser tip in the immediate vicinity of the support surface, e.g., the positions illustrated in Figs. 2B and 2C.

The desired deposition volume, i.e., bead volume, formed by this method is preferably in the range 2 pl (picoliters) to 2 nl (nanoliters), although volumes as high as 100 nl or more may be dispensed. It will be appreciated that the selected dispensed volume will depend on (i) the "footprint" of the dispenser tip, i.e., the size of the area spanned by the tip, (ii) the hydrophobicity of the support surface, and (iii) the time of contact with and rate of withdrawal of the tip from the support surface. In addition, bead size may be reduced by increasing the viscosity of the medium, effectively reducing the flow time of liquid from the dispenser onto the support surface. The drop size may be further constrained by depositing the drop in a hydrophilic region surrounded by a hydrophobic grid pattern on the support surface.

In a typical embodiment, the dispenser tip is tapped rapidly against the support surface, with a total residence time in contact with the support of less than about 1 msec, and a rate of upward travel from the surface of about 10 cm/sec.

Assuming that the bead that forms on contact with the surface is a hemispherical bead, with a diameter approximately equal to the width of the dispenser tip, as shown in Fig. 2C, the volume of the bead formed in relation to dispenser tip width (d) is given in Table 1 below. As seen, the volume of the bead ranges between 2 pl to 2 nl as the width size is increased from about 20 to 200 μm .

Table 1

d	Volume (nl)
20 μm	2×10^{-3}
50 μm	3.1×10^{-2}
100 μm	2.5×10^{-1}
200 μm	2

10 At a given tip size, bead volume can be reduced in
a controlled fashion by increasing surface
hydrophobicity, reducing time of contact of the tip
with the surface, increasing rate of movement of the
tip away from the surface, and/or increasing the
15 viscosity of the medium. Once these parameters are
fixed, a selected deposition volume in the desired pl
to nl range can be achieved in a repeatable fashion.

After depositing a bead at one selected location
on a support, the tip is typically moved to a
20 corresponding position on a second support, a droplet
is deposited at that position, and this process is
repeated until a liquid droplet of the reagent has been
deposited at a selected position on each of a plurality
of supports.

25 The tip is then washed to remove the reagent
liquid, filled with another reagent liquid and this
reagent is now deposited at each another array position
on each of the supports. In one embodiment, the tip is
washed and refilled by the steps of (i) dipping the
30 capillary channel of the device in a wash solution,
(ii) removing wash solution drawn into the capillary
channel, and (iii) dipping the capillary channel into
the new reagent solution.

From the foregoing, it will be appreciated that
35 the tweezers-like, open-capillary dispenser tip

provides the advantages that (i) the open channel of the tip facilitates rapid, efficient washing and drying before reloading the tip with a new reagent, (ii) passive capillary action can load the sample directly from a standard microwell plate while retaining sufficient sample in the open capillary reservoir for the printing of numerous arrays, (iii) open capillaries are less prone to clogging than closed capillaries, and (iv) open capillaries do not require a perfectly faced bottom surface for fluid delivery.

A portion of a microarray 36 formed on the surface 38 of a solid support 40 in accordance with the method just described is shown in Fig. 3. The array is formed of a plurality of analyte-specific reagent regions, such as regions 42, where each region may include a different analyte-specific reagent. As indicated above, the diameter of each region is preferably between about 20-200 μm . The spacing between each region and its closest (non-diagonal) neighbor, measured from center-to-center (indicated at 44), is preferably in the range of about 20-400 μm . Thus, for example, an array having a center-to-center spacing of about 250 μm contains about 40 regions/cm or 1,600 regions/cm². After formation of the array, the support is treated to evaporate the liquid of the droplet forming each region, to leave a desired array of dried, relatively flat regions. This drying may be done by heating or under vacuum.

In some cases, it is desired to first rehydrate the droplets containing the analyte reagents to allow for more time for adsorption to the solid support. It is also possible to spot out the analyte reagents in a humid environment so that droplets do not dry until the arraying operation is complete.

III. Automated Apparatus for Forming Arrays

In another aspect, the invention includes an automated apparatus for forming an array of analyte-assay regions on a solid support, where each region in the array has a known amount of a selected, analyte-specific reagent.

The apparatus is shown in planar, and partially schematic view in Fig. 4. A dispenser device 72 in the apparatus has the basic construction described above with respect to Fig. 1, and includes a dispenser 74 having an open-capillary channel terminating at a tip, substantially as shown in Figs. 1 and 2A-2C.

The dispenser is mounted in the device for movement toward and away from a dispensing position at which the tip of the dispenser taps a support surface, to dispense a selected volume of reagent solution, as described above. This movement is effected by a solenoid 76 as described above. Solenoid 76 is under the control of a control unit 77 whose operation will be described below. The solenoid is also referred to herein as dispensing means for moving the device into tapping engagement with a support, when the device is positioned at a defined array position with respect to that support.

The dispenser device is carried on an arm 74 which is threadedly mounted on a worm screw 80 driven (rotated) in a desired direction by a stepper motor 82 also under the control of unit 77. At its left end in the figure screw 80 is carried in a sleeve 84 for rotation about the screw axis. At its other end, the screw is mounted to the drive shaft of the stepper motor, which in turn is carried on a sleeve 86. The dispenser device, worm screw, the two sleeves mounting the worm screw, and the stepper motor used in moving the device in the "x" (horizontal) direction in the

figure form what is referred to here collectively as a displacement assembly 86.

The displacement assembly is constructed to produce precise, micro-range movement in the direction of the screw, i.e., along an x axis in the figure. In one mode, the assembly functions to move the dispenser in x-axis increments having a selected distance in the range 5-25 μm . In another mode, the dispenser unit may be moved in precise x-axis increments of several microns or more,; for positioning the dispenser at associated positions on adjacent supports, as will be described below.

The displacement assembly, in turn, is mounted for movement in the "y" (vertical) axis of the figure, for positioning the dispenser at a selected y axis position. The structure mounting the assembly includes a fixed rod 88 mounted rigidly between a pair of frame bars 90, 92, and a worm screw 94 mounted for rotation between a pair of frame bars 96, 98. The worm screw is driven (rotated) by a stepper motor 100 which operates under the control of unit 77. The motor is mounted on bar 96, as shown.

The structure just described, including worm screw 94 and motor 100, is constructed to produce precise, micro-range movement in the direction of the screw, i.e., along an y axis in the figure. As above, the structure functions in one mode to move the dispenser in y-axis increments having a selected distance in the range 5-250 μm , and in a second mode, to move the dispenser in precise y-axis increments of several microns (μm) or more, for positioning the dispenser at associated positions on adjacent supports.

The displacement assembly and structure for moving this assembly in the y axis are referred to herein collectively as positioning means for positioning the

dispensing device at a selected array position with respect to a support.

A holder 102 in the apparatus functions to hold a plurality of supports, such as supports 104 on which the microarrays of reagent regions are to be formed by the apparatus. The holder provides a number of recessed slots, such as slot 106, which receive the supports, and position them at precise selected positions with respect to the frame bars on which the dispenser moving means is mounted.

As noted above, the control unit in the device functions to actuate the two stepper motors and dispenser solenoid in a sequence designed for automated operation of the apparatus in forming a selected microarray of reagent regions on each of a plurality of supports.

The control unit is constructed, according to conventional microprocessor control principles, to provide appropriate signals to each of the solenoid and each of the stepper motors, in a given timed sequence and for appropriate signalling time. The construction of the unit, and the settings that are selected by the user to achieve a desired array pattern, will be understood from the following description of a typical apparatus operation.

Initially, one or more supports are placed in one or more slots in the holder. The dispenser is then moved to a position directly above a well (not shown) containing a solution of the first reagent to be dispensed on the support(s). The dispenser solenoid is actuated now to lower the dispenser tip into this well, causing the capillary channel in the dispenser to fill. Motors 82, 100 are now actuated to position the dispenser at a selected array position at the first of the supports. Solenoid actuation of the dispenser is

then effective to dispense a selected-volume droplet of that reagent at this location. As noted above, this operation is effective to dispense a selected volume preferably between 2 pl and 2 nl of the reagent solution.

The dispenser is now moved to the corresponding position at an adjacent support and a similar volume of the solution is dispensed at this position. The process is repeated until the reagent has been dispensed at this preselected corresponding position on each of the supports.

Where it is desired to dispense a single reagent at more than two array positions on a support, the dispenser may be moved to different array positions at each support, before moving the dispenser to a new support, or solution can be dispensed at individual positions on each support, at one selected position, then the cycle repeated for each new array position.

To dispense the next reagent, the dispenser is positioned over a wash solution (not shown), and the dispenser tip is dipped in and out of this solution until the reagent solution has been substantially washed from the tip. Solution can be removed from the tip, after each dipping, by vacuum, compressed air spray, sponge, or the like.

The dispenser tip is now dipped in a second reagent well, and the filled tip is moved to a second selected array position in the first support. The process of dispensing reagent at each of the corresponding second-array positions is then carried as above. This process is repeated until an entire microarray of reagent solutions on each of the supports has been formed.

IV. Microarray Substrate

This section describes embodiments of a substrate having a microarray of biological polymers carried on the substrate surface. Subsection A describes a multi-cell substrate, each cell of which contains a microarray, and preferably an identical microarray, of distinct biopolymers, such as distinct polynucleotides, formed on a porous surface. Subsection B describes a microarray of distinct polynucleotides bound on a glass slide coated with a polycationic polymer.

A. Multi-Cell Substrate

Fig. 9 illustrates, in plan view, a substrate constructed according to the invention. The substrate has an 8 x 12 rectangular array 112 of cells, such as cells 114, 116, formed on the substrate surface. With reference to Fig. 10, each cell, such as cell 114, in turn supports a microarray 118 of distinct biopolymers, such as polypeptides or polynucleotides at known, addressable regions of the microarray. Two such regions forming the microarray are indicated at 120, and correspond to regions, such as regions 42, forming the microarray of distinct biopolymers shown in Fig. 3.

The 96-cell array shown in Fig. 9 has typically array dimensions between about 12 and 244 mm in width and 8 and 400 mm in length, with the cells in the array having width and length dimension of 1/12 and 1/8 the array width and length dimensions, respectively, i.e., between about 1 and 20 in width and 1 and 50 mm in length.

The construction of substrate is shown cross-sectionally in Fig. 11, which is an enlarged sectional view taken along view line 124 in Fig. 9. The substrate includes a water-impermeable backing 126, such as a glass slide or rigid polymer sheet. Formed on the surface of the backing is a water-permeable film

128. The film is formed of a porous membrane material, such as nitrocellulose membrane, or a porous web material, such as a nylon, polypropylene, or PVDF porous polymer material. The thickness of the film is preferably between about 10 and 1000 μm . The film may be applied to the backing by spraying or coating uncured material on the backing, or by applying a preformed membrane to the backing. The backing and film may be obtained as a preformed unit from commercial source, e.g., a plastic-backed nitrocellulose film available from Schleicher and Schuell Corporation.

With continued reference to Fig. 11, the film-covered surface in the substrate is partitioned into a desired array of cells by water-impermeable grid lines, such as lines 130, 132, which have infiltrated the film down to the level of the backing, and extend above the surface of the film as shown, typically a distance of 100 to 2000 μm above the film surface.

The grid lines are formed on the substrate by laying down an uncured or otherwise flowable resin or elastomer solution in an array grid, allowing the material to infiltrate the porous film down to the backing, then curing or otherwise hardening the grid lines to form the cell-array substrate.

One preferred material for the grid is a flowable silicone available from Loctite Corporation. The barrier material can be extruded through a narrow syringe (e.g., 22 gauge) using air pressure or mechanical pressure. The syringe is moved relative to the solid support to print the barrier elements as a grid pattern. The extruded bead of silicone wicks into the pores of the solid support and cures to form a shallow waterproof barrier separating the regions of the solid support.

In alternative embodiments, the barrier element can be a wax-based material or a thermoset material such as epoxy. The barrier material can also be a UV-curing polymer which is exposed to UV light after being printed onto the solid support. The barrier material may also be applied to the solid support using printing techniques such as silk-screen printing. The barrier material may also be a heat-seal stamping of the porous solid support which seals its pores and forms a water-impervious barrier element. The barrier material may also be a shallow grid which is laminated or otherwise adhered to the solid support.

In addition to plastic-backed nitrocellulose, the solid support can be virtually any porous membrane with or without a non-porous backing. Such membranes are readily available from numerous vendors and are made from nylon, PVDF, polysulfone and the like. In an alternative embodiment, the barrier element may also be used to adhere the porous membrane to a non-porous backing in addition to functioning as a barrier to prevent cross contamination of the assay reagents.

In an alternative embodiment, the solid support can be of a non-porous material. The barrier can be printed either before or after the microarray of biomolecules is printed on the solid support.

As can be appreciated, the cells formed by the grid lines and the underlying backing are water-impermeable, having side barriers projecting above the porous film in the cells. Thus, defined-volume samples can be placed in each well without risk of cross-contamination with sample material in adjacent cells. In Fig. 11, defined volume samples, such as sample 134, are shown in the cells.

As noted above, each well contains a microarray of distinct biopolymers. In one general embodiment, the

microarrays in the well are identical arrays of distinct biopolymers, e.g., different sequence polynucleotides. Such arrays can be formed in accordance with the methods described in Section II, by
5 depositing a first selected polynucleotide at the same selected microarray position in each of the cells, then depositing a second polynucleotide at a different microarray position in each well, and so on until a complete, identical microarray is formed in each cell.

10 In a preferred embodiment, each microarray contains about 10^3 distinct polynucleotide or polypeptide biopolymers per surface area of less than about 1 cm^2 . Also in a preferred embodiment, the biopolymers in each microarray region are present in a
15 defined amount between about 0.1 femtomoles and 100 nanomoles. The ability to form high-density arrays of biopolymers, where each region is formed of a well-defined amount of deposited material, can be achieved in accordance with the microarray-forming method
20 described in Section II.

Also in a preferred embodiments, the biopolymers are polynucleotides having lengths of at least about 50 bp, i.e., substantially longer than oligonucleotides which can be formed in high-density arrays by schemes
25 involving parallel, step-wise polymer synthesis on the array surface.

In the case of a polynucleotide array, in an assay procedure, a small volume of the labeled DNA probe mixture in a standard hybridization solution is loaded
30 onto each cell. The solution will spread to cover the entire microarray and stop at the barrier elements. The solid support is then incubated in a humid chamber at the appropriate temperature as required by the assay.

Each assay may be conducted in an "open-face" format where no further sealing step is required, since the hybridization solution will be kept properly hydrated by the water vapor in the humid chamber. At the conclusion of the incubation step, the entire solid support containing the numerous microarrays is rinsed quickly enough to dilute the assay reagents so that no significant cross contamination occurs. The entire solid support is then reacted with detection reagents if needed and analyzed using standard colorimetric, radioactive or fluorescent detection means. All processing and detection steps are performed simultaneously to all of the microarrays on the solid support ensuring uniform assay conditions for all of the microarrays on the solid support.

B. Glass-Slide Polynucleotide Array

Fig. 5 shows a substrate 136 formed according to another aspect of the invention, and intended for use in detecting binding of labeled polynucleotides to one or more of a plurality distinct polynucleotides. The substrate includes a glass substrate 138 having formed on its surface, a coating of a polycationic polymer, preferably a cationic polypeptide, such as polylysine or polyarginine. Formed on the polycationic coating is a microarray 140 of distinct polynucleotides, each localized at known selected array regions, such as regions 142.

The slide is coated by placing a uniform-thickness film of a polycationic polymer, e.g., poly-l-lysine, on the surface of a slide and drying the film to form a dried coating. The amount of polycationic polymer added is sufficient to form at least a monolayer of polymers on the glass surface. The polymer film is bound to surface via electrostatic binding between

negative silyl-OH groups on the surface and charged amine groups in the polymers. Poly-L-lysine coated glass slides may be obtained commercially, e.g., from Sigma Chemical Co. (St. Louis, MO).

5 To form the microarray, defined volumes of distinct polynucleotides are deposited on the polymer-coated slide, as described in Section II. According to an important feature of the substrate, the deposited polynucleotides remain bound to the coated slide
10 surface non-covalently when an aqueous DNA sample is applied to the substrate under conditions which allow hybridization of reporter-labeled polynucleotides in the sample to complementary-sequence (single-stranded) polynucleotides in the substrate array. The method is
15 illustrated in Examples 1 and 2.

 To illustrate this feature, a substrate of the type just described, but having an array of same-sequence polynucleotides, was mixed with fluorescent-labeled complementary DNA under hybridization
20 conditions. After washing to remove non-hybridized material, the substrate was examined by low-power fluorescence microscopy. The array can be visualized by the relatively uniform labeling pattern of the array regions.

25 In a preferred embodiment, each microarray contains at least 10^3 distinct polynucleotide or polypeptide biopolymers per surface area of less than about 1 cm^2 . In the embodiment shown in Fig. 5, the microarray contains 400 regions in an area of about 16 mm^2 , or 2.5×10^3 regions/ cm^2 . Also in a preferred
30 embodiment, the polynucleotides in the each microarray region are present in a defined amount between about 0.1 femtomoles and 100 nanomoles in the case of polynucleotides. As above, the ability to form high-

density arrays of this type, where each region is formed of a well-defined amount of deposited material, can be achieved in accordance with the microarray-forming method described in Section II.

5 Also in a preferred embodiments, the polynucleotides have lengths of at least about 50 bp, i.e., substantially longer than oligonucleotides which can be formed in high-density arrays by various in situ synthesis schemes.

10

V. Utility

Microarrays of immobilized nucleic acid sequences prepared in accordance with the invention can be used for large scale hybridization assays in numerous
15 genetic applications, including genetic and physical mapping of genomes, monitoring of gene expression, DNA sequencing, genetic diagnosis, genotyping of organisms, and distribution of DNA reagents to researchers.

For gene mapping, a gene or a cloned DNA fragment
20 is hybridized to an ordered array of DNA fragments, and the identity of the DNA elements applied to the array is unambiguously established by the pixel or pattern of pixels of the array that are detected. One application of such arrays for creating a genetic map is described
25 by Nelson, et al. (1993). In constructing physical maps of the genome, arrays of immobilized cloned DNA fragments are hybridized with other cloned DNA fragments to establish whether the cloned fragments in the probe mixture overlap and are therefore contiguous
30 to the immobilized clones on the array. For example, Lehrach, et al., describe such a process.

The arrays of immobilized DNA fragments may also be used for genetic diagnostics. To illustrate, an array containing multiple forms of a mutated gene or
35 genes can be probed with a labeled mixture of a

patient's DNA which will preferentially interact with only one of the immobilized versions of the gene.

The detection of this interaction can lead to a medical diagnosis. Arrays of immobilized DNA fragments can also be used in DNA probe diagnostics. For example, the identity of a pathogenic microorganism can be established unambiguously by hybridizing a sample of the unknown pathogen's DNA to an array containing many types of known pathogenic DNA. A similar technique can also be used for unambiguous genotyping of any organism. Other molecules of genetic interest, such as cDNA's and RNA's can be immobilized on the array or alternately used as the labeled probe mixture that is applied to the array.

In one application, an array of cDNA clones representing genes is hybridized with total cDNA from an organism to monitor gene expression for research or diagnostic purposes. Labeling total cDNA from a normal cell with one color fluorophore and total cDNA from a diseased cell with another color fluorophore and simultaneously hybridizing the two cDNA samples to the same array of cDNA clones allows for differential gene expression to be measured as the ratio of the two fluorophore intensities. This two-color experiment can be used to monitor gene expression in different tissue types, disease states, response to drugs, or response to environmental factors. & An example of this approach is illustrated in Examples 2, described with respect to Fig. 8.

By way of example and without implying a limitation of scope, such a procedure could be used to simultaneously screen many patients against all known mutations in a disease gene. This invention could be used in the form of, for example, 96 identical 0.9 cm x 2.2 cm microarrays fabricated on a single 12 cm x 18 cm

sheet of plastic-backed nitrocellulose where each microarray could contain, for example, 100 DNA fragments representing all known mutations of a given gene. The region of interest from each of the DNA samples from 96 patients could be amplified, labeled, and hybridized to the 96 individual arrays with each assay performed in 100 microliters of hybridization solution. The approximately 1 thick silicone rubber barrier elements between individual arrays prevent cross contamination of the patient samples by sealing the pores of the nitrocellulose and by acting as a physical barrier between each microarray. The solid support containing all 96 microarrays assayed with the 96 patient samples is incubated, rinsed, detected and analyzed as a single sheet of material using standard radioactive, fluorescent, or colorimetric detection means (Maniatis, et al., 1989). Previously, such a procedure would involve the handling, processing and tracking of 96 separate membranes in 96 separate sealed chambers. By processing all 96 arrays as a single sheet of material, significant time and cost savings are possible.

The assay format can be reversed where the patient or organism's DNA is immobilized as the array elements and each array is hybridized with a different mutated allele or genetic marker. The gridded solid support can also be used for parallel non-DNA ELISA assays. Furthermore, the invention allows for the use of all standard detection methods without the need to remove the shallow barrier elements to carry out the detection step.

In addition to the genetic applications listed above, arrays of whole cells, peptides, enzymes, antibodies, antigens, receptors, ligands, phospholipids, polymers, drug cogener preparations or

chemical substances can be fabricated by the means described in this invention for large scale screening assays in medical diagnostics, drug discovery, molecular biology, immunology and toxicology.

5 The multi-cell substrate aspect of the invention allows for the rapid and convenient screening of many DNA probes against many ordered arrays of DNA fragments. This eliminates the need to handle and detect many individual arrays for performing mass
10 screenings for genetic research and diagnostic applications. Numerous microarrays can be fabricated on the same solid support and each microarray reacted with a different DNA probe while the solid support is processed as a single sheet of material.

15

The following examples illustrate, but in no way are intended to limit, the present invention.

Example 1

20 Genomic-Complexity Hybridization to Micro
DNA Arrays Representing the Yeast
Saccharomyces cerevisiae Genome with
Two-Color Fluorescent Detection

The array elements were randomly amplified PCR
25 (Bohlander, et al., 1992) products using physically mapped lambda clones of *S. cerevisiae* genomic DNA templates (Riles, et al., 1993). The PCR was performed directly on the lambda phage lysates resulting in an amplification of both the 35 kb lambda vector and the
30 5-15 kb yeast insert sequences in the form of a uniform distribution of PCR product between 250-1500 base pairs in length. The PCR product was purified using Sephadex G50 gel filtration (Pharmacia, Piscataway, NJ) and concentrated by evaporation to dryness at room
35 temperature overnight. Each of the 864 amplified

lambda clones was rehydrated in 15 μ l of 3 \times SSC in preparation for spotting onto the glass.

The micro arrays were fabricated on microscope slides which were coated with a layer of poly-l-lysine (Sigma). The automated apparatus described in Section IV loaded 1 μ l of the concentrated lambda clone PCR product in 3 \times SSC directly from 96 well storage plates into the open capillary printing element and deposited -5 nl of sample per slide at 380 micron spacing between spots, on each of 40 slides. The process was repeated for all 864 samples and 8 control spots. After the spotting operation was complete, the slides were rehydrated in a humid chamber for 2 hours, baked in a dry 80° vacuum oven for 2 hours, rinsed to remove unabsorbed DNA and then treated with succinic anhydride to reduce non-specific adsorption of the labeled hybridization probe to the poly-l-lysine coated glass surface. Immediately prior to use, the immobilized DNA on the array was denatured in distilled water at 90° for 2 minutes.

For the pooled chromosome experiment, the 16 chromosomes of *Saccharomyces cerevisiae* were separated in a CHEF agarose gel apparatus (Biorad, Richmond, CA). The six largest chromosomes were isolated in one gel slice and the smallest 10 chromosomes in a second gel slice. The DNA was recovered using a gel extraction kit (Qiagen, Chatsworth, CA). The two chromosome pools were randomly amplified in a manner similar to that used for the target lambda clones. Following amplification, 5 micrograms of each of the amplified chromosome pools were separately random-primer labeled using Klenow polymerase (Amersham, Arlington Heights, IL) with a lissamine conjugated nucleotide analog (Dupont NEN, Boston, MA) for the pool containing the six largest chromosomes, and with a fluorescein

conjugated nucleotide analog (BMB) for the pool containing smallest ten chromosomes. The two pools were mixed and concentrated using an ultrafiltration device (Amicon, Danvers, MA).

5 Five micrograms of the hybridization probe consisting of both chromosome pools in 7.5 μ l of TE was denatured in a boiling water bath and then snap cooled on ice. 2.5 μ l of concentrated hybridization solution (5 \times SSC and 0.1% SDS) was added and all 10 μ l
10 transferred to the array surface, covered with a cover slip, placed in a custom-built single-slide humidity chamber and incubated at 60° for 12 hours. The slides were then rinsed at room temperature in 0.1 \times SSC and 0.1%SDS for 5 minutes, cover slipped and scanned.

15 A custom built laser fluorescent scanner was used to detect the two-color hybridization signals from the 1.8 \times 1.8 cm array at 20 micron resolution. The scanned image was gridded and analyzed using custom image analysis software. After correcting for optical
20 crosstalk between the fluorophores due to their overlapping emission spectra, the red and green hybridization values for each clone on the array were correlated to the known physical map position of the clone resulting in a computer-generated color karyotype
25 of the yeast genome.

Figure 6 shows the hybridization pattern of the two chromosome pools. A red signal indicates that the lambda clone on the array surface contains a cloned genomic DNA segment from one of the largest six yeast
30 chromosomes. A green signal indicates that the lambda clone insert comes from one of the smallest ten yeast chromosomes. Orange signals indicate repetitive sequences which cross hybridized to both chromosome pools. Control spots on the array confirm that the
35 hybridization is specific and reproducible.

The physical map locations of the genomic DNA fragments contained in each of the clones used as array elements have been previously determined by Olson and co-workers (Riles, et al.) allowing for the automatic generation of the color karyotype shown in Figure 7. The color of a chromosomal section on the karyotype corresponds to the color of the array element containing the clone from that section. The black regions of the karyotype represent false negative dark spots on the array (10%) or regions of the genome not covered by the Olson clone library (90%). Note that the largest six chromosomes are mainly red while the smallest ten chromosomes are mainly green matching the original CHEF gel isolation of the hybridization probe. Areas of the red chromosomes containing green spots and vice-versa are probably due to spurious sample tracking errors in the formation of the original library and in the amplification and spotting procedures.

The yeast genome arrays have also been probed with individual clones or pools of clones that are fluorescently labeled for physical mapping purposes. The hybridization signals of these clones to the array were translated into a position on the physical map of yeast.

25

Example 2

Total cDNA Hybridized to Micro Arrays of cDNA Clones with Two-Color Fluorescent Detection

24 clones containing cDNA inserts from the plant *Arabidopsis* were amplified using PCR. Salt was added to the purified PCR products to a final concentration of 3 x SSC. The cDNA clones were spotted on poly-l-lysine coated microscope slides in a manner similar to Example 1. Among the cDNA clones was a clone

35

representing a transcription factor HAT 4, which had previously been used to create a transgenic line of the plant *Arabidopsis*, in which this gene is present at ten times the level found in wild-type *Arabidopsis* (Schena, et al., 1992).

Total poly-A mRNA from wild type *Arabidopsis* was isolated using standard methods (Maniatis, et al., 1989) and reverse transcribed into total cDNA, using fluorescein nucleotide analog to label the cDNA product (green fluorescence). A similar procedure was performed with the transgenic line of *Arabidopsis* where the transcription factor HAT4 was inserted into the genome using standard gene transfer protocols. cDNA copies of mRNA from the transgenic plant are labeled with a lissamine nucleotide analog (red fluorescence). Two micrograms of the cDNA products from each type of plant were pooled together and hybridized to the cDNA clone array in a 10 microliter hybridization reaction in a manner similar to Example 1. Rinsing and detection of hybridization was also performed in a manner similar to Example 1. Fig. 8 show the resulting hybridization pattern of the array.

Genes equally expressed in wild type and the transgenic *Arabidopsis* appeared yellow due to equal contributions of the green and red fluorescence to the final signal. The dots are different intensities of yellow indicating various levels of gene expression. The cDNA clone representing the transcription factor HAT4, expressed in the transgenic line of *Arabidopsis* but not detectably expressed in wild type *Arabidopsis*, appears as a red dot (with the arrow pointing to it), indicating the preferential expression of the transcription factor in the red-labeled transgenic *Arabidopsis* and the relative lack of expression of the

transcription factor in the green-labeled wild type *Arabidopsis*.

An advantage of the microarray hybridization format for gene expression studies is the high partial concentration of each cDNA species achievable in the 10 microliter hybridization reaction. This high partial concentration allows for detection of rare transcripts without the need for PCR amplification of the hybridization probe which may bias the true genetic representation of each discrete cDNA species.

Gene expression studies such as these can be used for genomics research to discover which genes are expressed in which cell types, disease states, development states or environmental conditions. Gene expression studies can also be used for diagnosis of disease by empirically correlating gene expression patterns to disease states.

Example 3

Multiplexed Colorimetric Hybridization on a Gridded Solid Support

A sheet of plastic-backed nitrocellulose was gridded with barrier elements made from silicone rubber according to the description in Section IV-A. The sheet was soaked in 10 × SSC and allowed to dry. As shown in Fig. 12, 192 M13 clones each with a different yeast inserts were arrayed 400 microns apart in four quadrants of the solid support using the automated device described in Section III. The bottom left quadrant served as a negative control for hybridization while each of the other three quadrants was hybridized simultaneously with a different oligonucleotide using the open-face hybridization technology described in Section IV-A. The first two and last four elements of

each array are positive controls for the colorimetric detection step.

5 The oligonucleotides were labeled with fluorescein which was detected using an anti-fluorescein antibody conjugated to alkaline phosphatase that precipitated an NBT/BCIP dye on the solid support (Amersham). Perfect matches between the labeled oligos and the M13 clones resulted in dark spots visible to the naked eye and detected using an optical scanner (HP ScanJet II) attached to a personal computer. The hybridization patterns are different in every quadrant indicating that each oligo found several unique M13 clones from among the 192 with a perfect sequence match. Note that the open capillary printing tip leaves detectable
10 dimples on the nitrocellulose which can be used to
15 automatically align and analyze the images.

Although the invention has been described with respect to specific embodiments and methods, it will be
20 clear that various changes and modification may be made without departing from the invention.

IT IS CLAIMED:

1. A method of forming a microarray of analyte-assay regions on a solid support, where each region in the array has a known amount of a selected, analyte-specific reagent, said method comprising,
- 5 (a) loading a solution of a selected analyte-specific reagent in a reagent-dispensing device having an elongate capillary channel (i) formed by spaced-apart, coextensive elongate members, (ii) adapted to hold a quantity of the reagent solution and (iii) having a tip region at which aqueous solution in the channel forms a meniscus,
- 10 (b) tapping the tip of the dispensing device against a solid support at a defined position on the surface, with an impulse effective to break the meniscus in the capillary channel and deposit a selected volume of solution on the surface, and
- 15 (c) repeating steps (a) and (b) until said array is formed.
- 20
2. The method of claim 1, wherein said tapping is carried out with an impulse effective to deposit a selected volume in the volume range between 0.01 to 100 nl.
- 25
3. The method of claim 1, wherein said channel is formed by a pair of spaced-apart tapered elements.
- 30
4. The method of claim 1, for forming a plurality of such arrays, wherein step (b) is applied to a selected position on each of a plurality of solid supports at each repeat cycle proceeding step (c).

5. The method of claim 1, which further includes, after performing steps (a) and (b) at least one time, reloading the reagent-dispensing device with a new reagent solution by the steps of (i) dipping the capillary channel of the device in a wash solution, (ii) removing wash solution drawn into the capillary channel, and (iii) dipping the capillary channel into the new reagent solution.

6. Automated apparatus for forming a microarray of analyte-assay regions on a plurality of solid supports, where each region in the array has a known amount of a selected, analyte-specific reagent, said apparatus comprising

(a) a holder for holding, at known positions, a plurality of planar supports,

(b) a reagent dispensing device having an open capillary channel (i) formed by spaced-apart, coextensive elongate members (ii) adapted to hold a quantity of the reagent solution and (iii) having a tip region at which aqueous solution in the channel forms a meniscus,

(c) positioning means for positioning the dispensing device at a selected array position with respect to a support in said holder,

(d) dispensing means for moving the device into tapping engagement against a support with a selected impulse, when the device is positioned at a defined array position with respect to that support, with an impulse effective to break the meniscus of liquid in the capillary channel and deposit a selected volume of solution on the surface, and

(e) control means for controlling said positioning and dispensing means.

7. The apparatus of claim 6, wherein said dispensing means is effective to move said dispensing device against a support with an impulse effective to deposit a selected volume in the volume range between
5 0.01 to 100 nl.

8. The apparatus of claim 6, wherein said channel is formed by a pair of spaced-apart tapered elements.

10 9. The apparatus of claim 6, wherein the control means operates to (i) place the dispensing device at a loading station, (ii) move the capillary channel in the device into a selected reagent at the loading station, to load the dispensing device with the reagent, and
15 (iii) dispense the reagent at a defined array position on each of the supports on said holder.

10. The apparatus of claim 6, wherein the control device further operates, at the end of a dispensing
20 cycle, to wash the dispensing device by (i) placing the dispensing device at a washing station, (ii) moving the capillary channel in the device into a wash fluid, to load the dispensing device with the fluid, and (iii) remove the wash fluid prior to loading the dispensing
25 device with a fresh selected reagent.

11. The apparatus of claim 6, wherein said device is one of a plurality of such devices which are carried on the arm for dispensing different analyte assay
30 reagents at selected spaced array positions.

12. A substrate with a surface having a microarray of at least 10^3 distinct polynucleotide or polypeptide biopolymers per 1 cm^2 surface area, each

distinct biopolymer sample (i) being disposed at a separate, defined position in said array, (ii) having a length of at least 50 subunits, and (iii) being present in a defined amount between about 0.1 femtomole and 100 nanomoles.

13. The substrate of claim 12, wherein said surface is glass slide coated with polylysine, and said biopolymers are polynucleotides.

14. The substrate of claim 12, wherein said substrate has a water-impermeable backing, a water-permeable film formed on the backing, and a grid formed on the film, where said grid (i) is composed of intersecting water-impervious grid elements extending from said backing to positions raised above the surface of said film, and (ii) partitions the film into a plurality of water-impervious cells, where each cell contains such a biopolymer array.

15. A substrate with a surface array of sample-receiving cells, comprising
a water-impermeable backing,
a water-permeable film formed on the backing, and
a grid formed on the film, said grid being composed of intersecting water-impervious grid elements extending from said backing to positions raised above the surface of said film.

16. The substrate of claim 15, wherein the cells of the array each contain an array of biopolymers.

17. A substrate for use in detecting binding of labeled biopolymers to one or more of a plurality distinct polynucleotides, comprising

a non-porous, glass substrate,
a coating of a cationic polymer on said substrate,
and

an array of distinct polynucleotides to said
5 coating, where each biopolymer is disposed at a
separate, defined position in a surface array of
biopolymers.

18. A method of detecting differential expression
10 of each of a plurality of genes in a first cell type
with respect to expression of the same genes in a
second cell types, said method comprising

producing fluorescence-labeled cDNA's from mRNA's
isolated from the two cells types, where the cDNA's
15 from the first and second cells are labeled with first
and second different fluorescent reporters,

adding a mixture of the labeled cDNA's from the
two cell types to an array of polynucleotides
representing a plurality of known genes derived from
20 the two cell types, under conditions that result in
hybridization of the cDNA's to complementary-sequence
polynucleotides in the array; and

examining the array by fluorescence under
fluorescence excitation conditions in which (i)
25 polynucleotides in the array that are hybridized
predominantly to cDNA's derived from one of the first
and second cell types give a distinct first or second
fluorescence emission color, respectively, and (ii)
polynucleotides in the array that are hybridized to
30 substantially equal numbers of cDNA's derived from the
first and second cell types give a distinct combined
fluorescence emission color, respectively,

wherein the relative expression of known genes in
the two cell types can be determined by the observed
35 fluorescence emission color of each spot.

19. The method of claim 18, wherein the array of polynucleotides is formed on a substrate with a surface having an array of at least 10^2 distinct polynucleotide or polypeptide biopolymers in a surface area of less than about 1 cm^2 , each distinct biopolymer (i) being disposed at a separate, defined position in said array, (ii) having a length of at least 50 subunits, and (iii) being present in a defined amount between about .1 femtomole and 100 nmoles.

10

20. The method of claim 19, wherein said surface is a glass slide coated with polylysine, and said biopolymers are polynucleotides non-covalently bound to said polylysine.

15

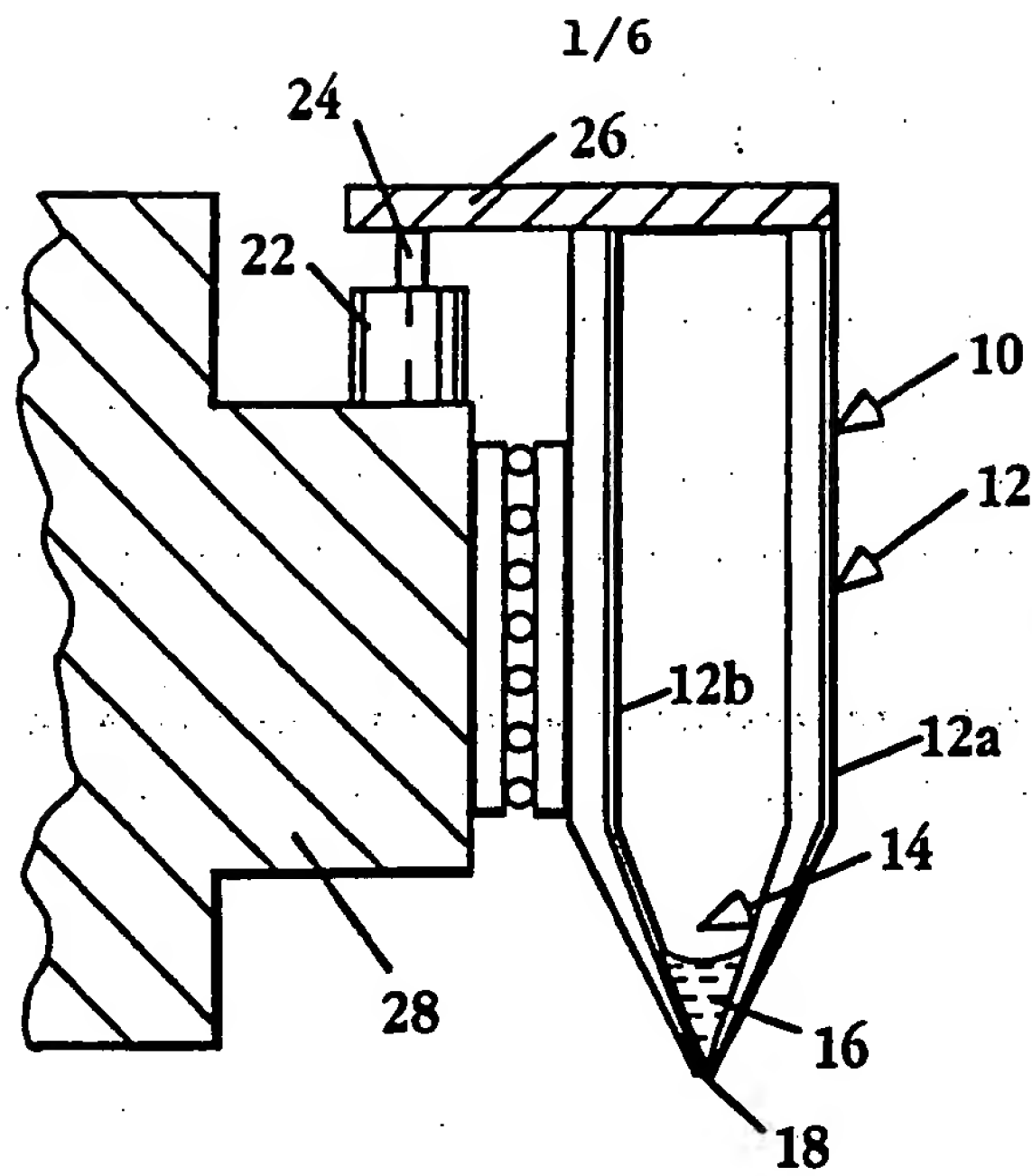


Fig. 1

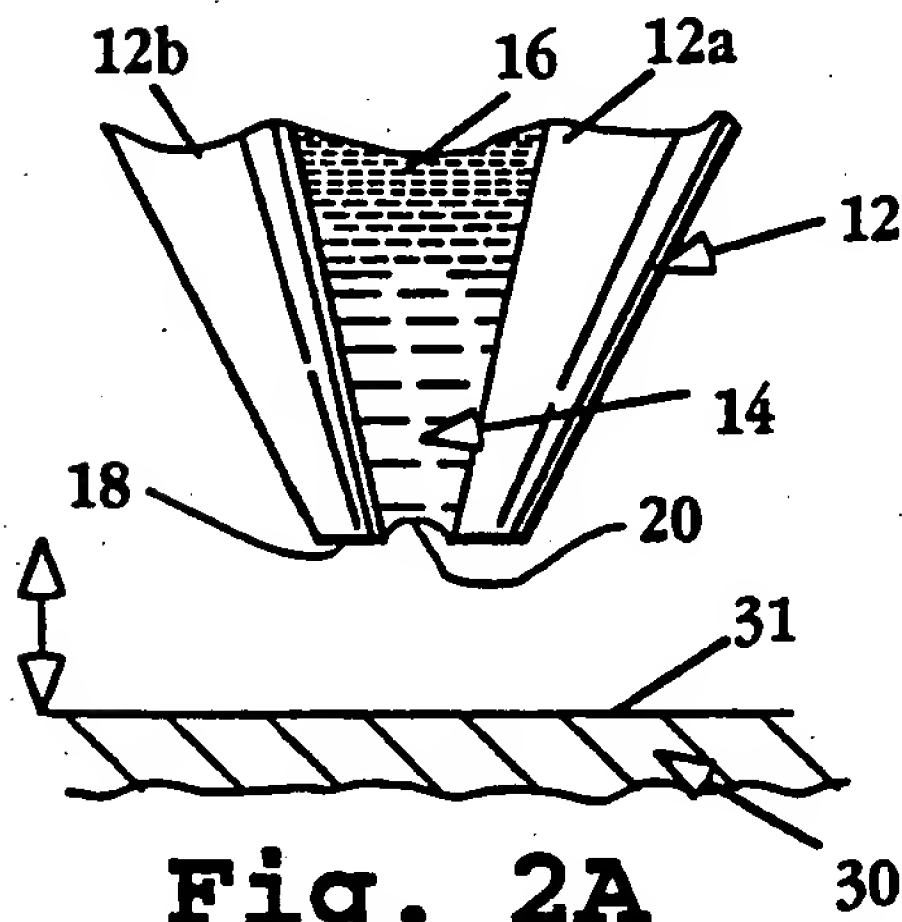


Fig. 2A

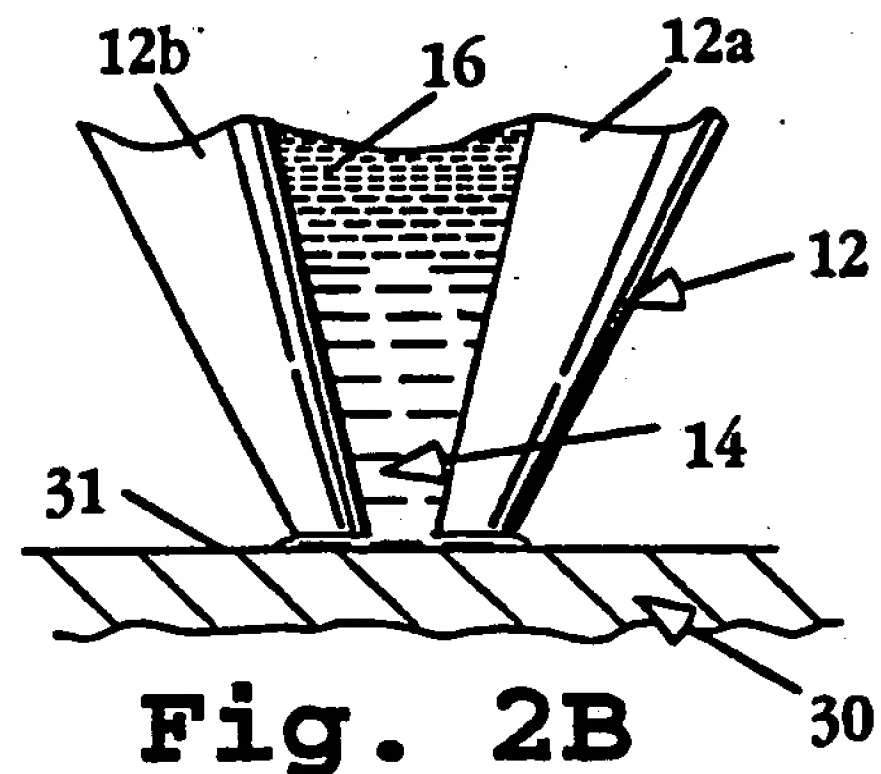


Fig. 2B

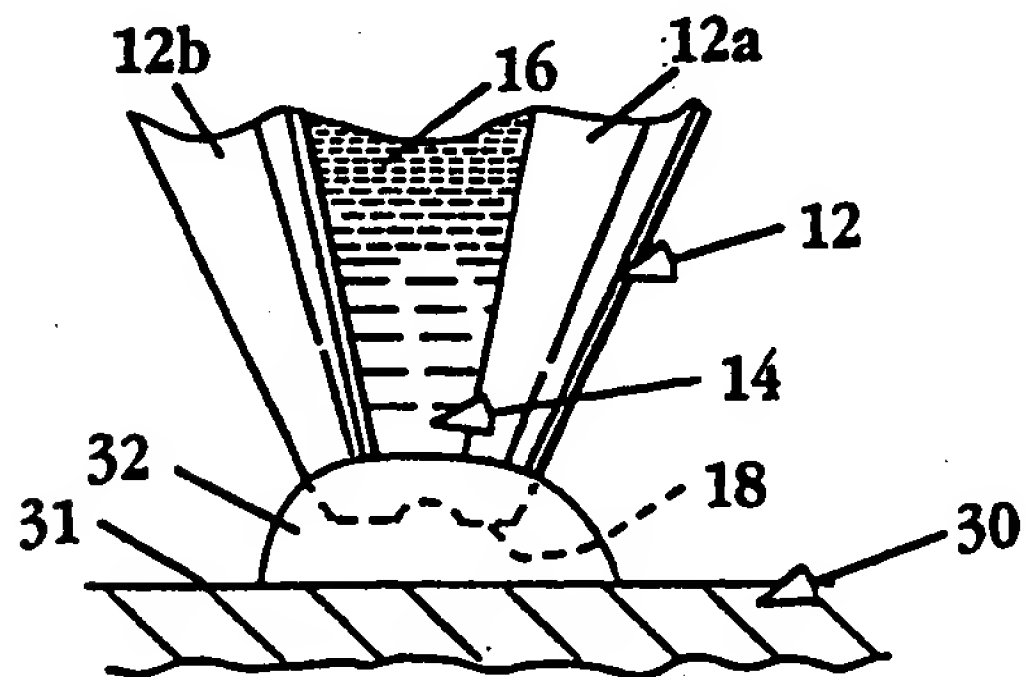


Fig. 2C

2/6

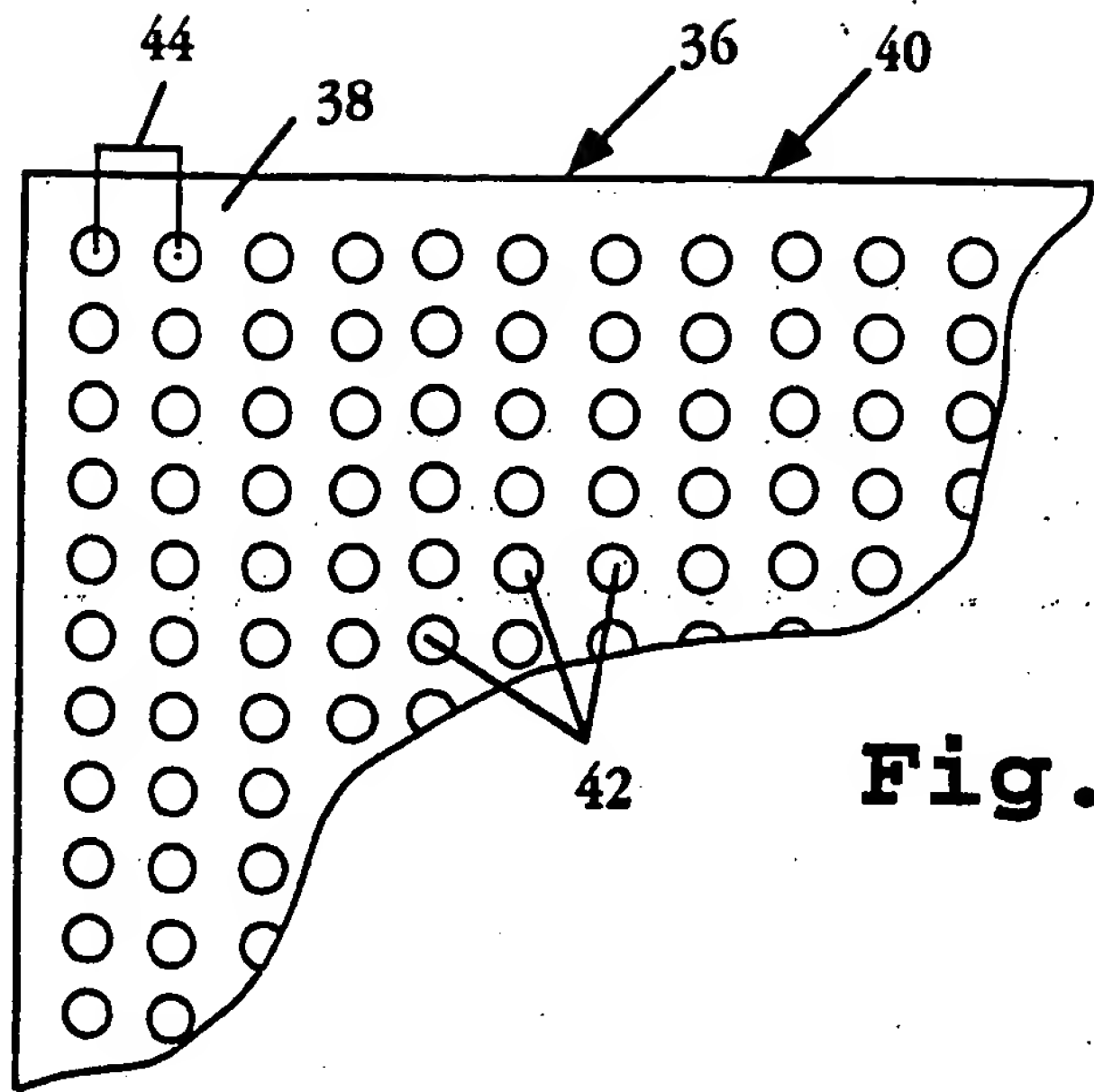


Fig. 3

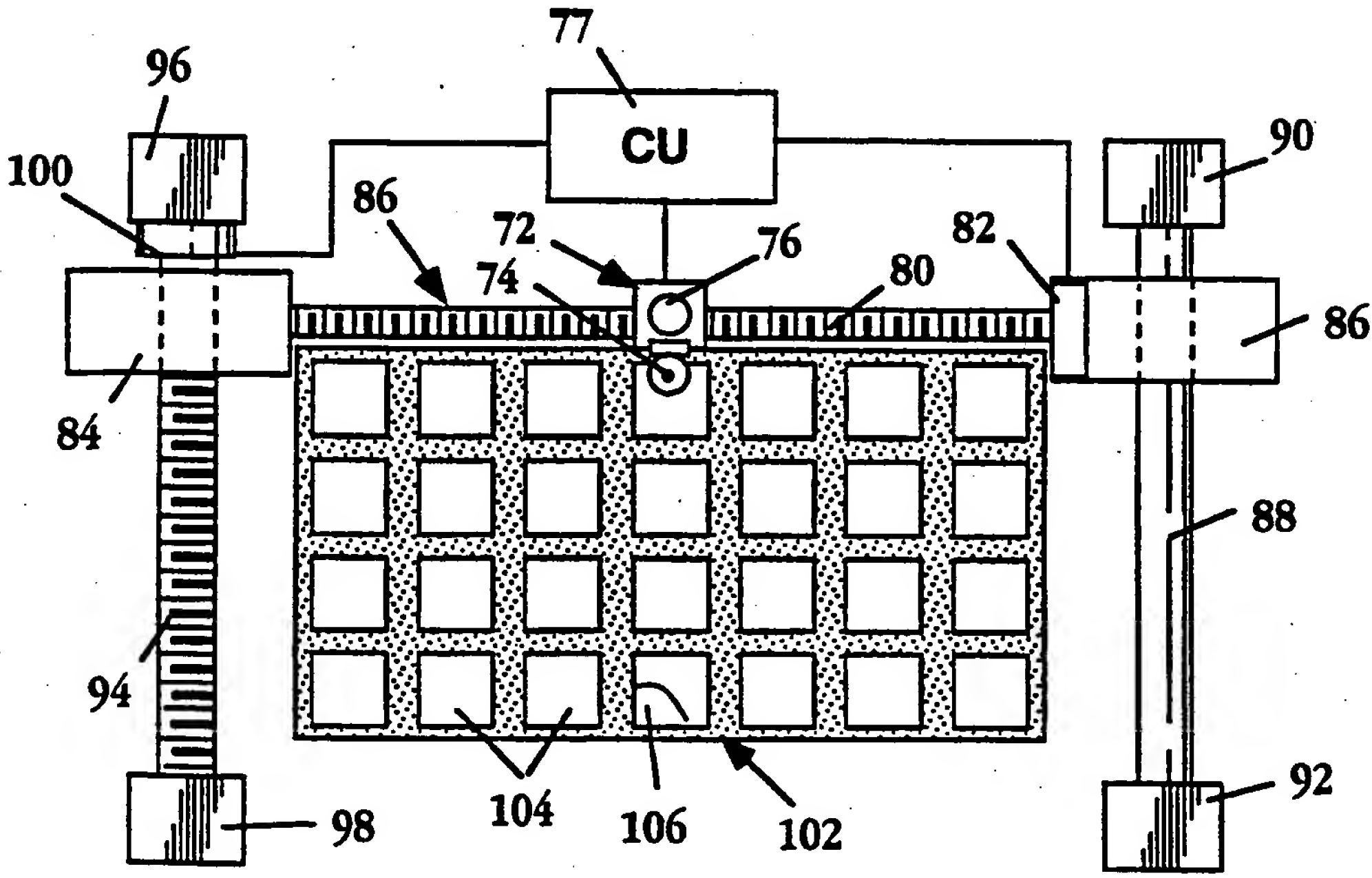


Fig. 4

3/6

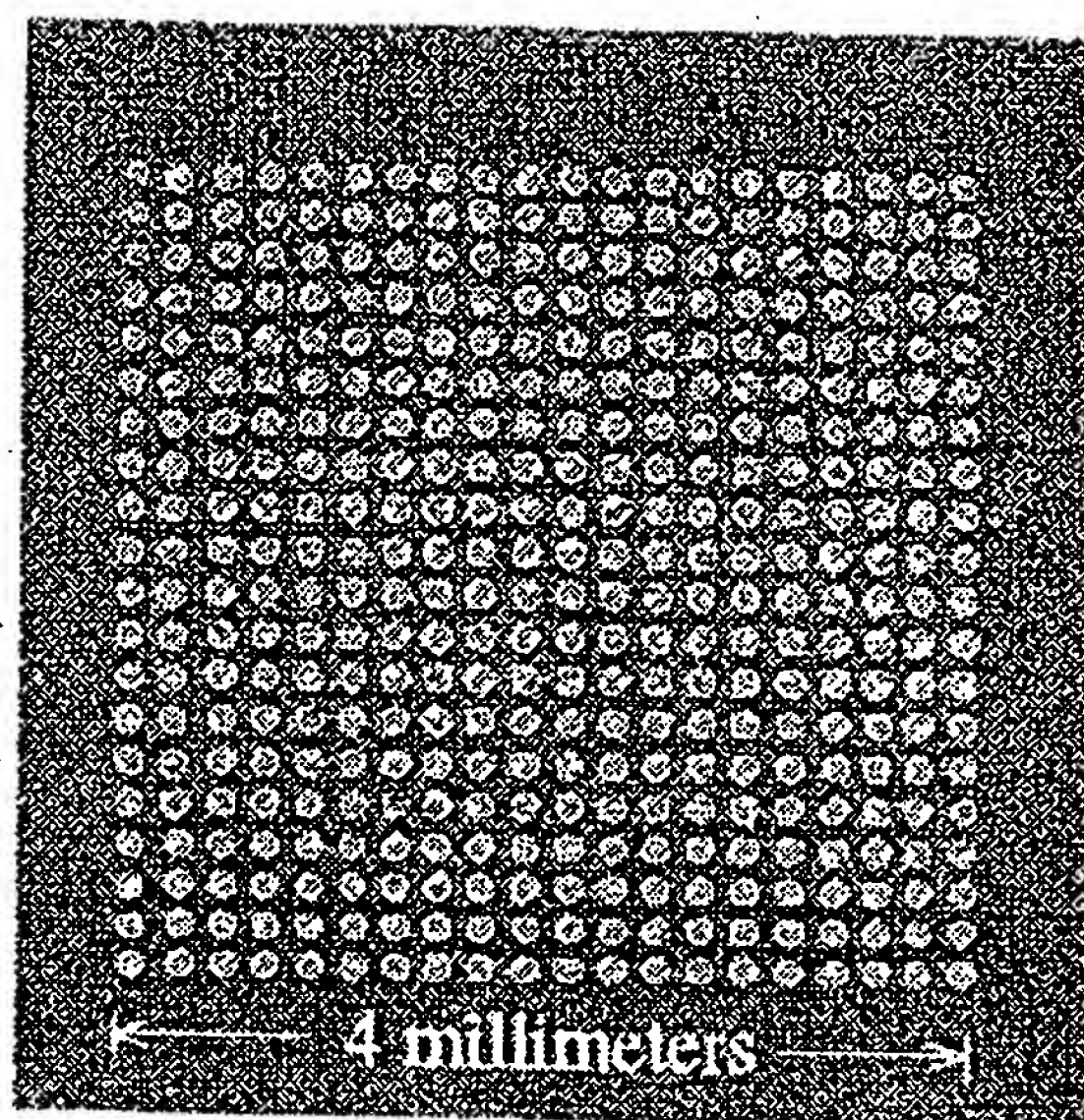


Fig. 5

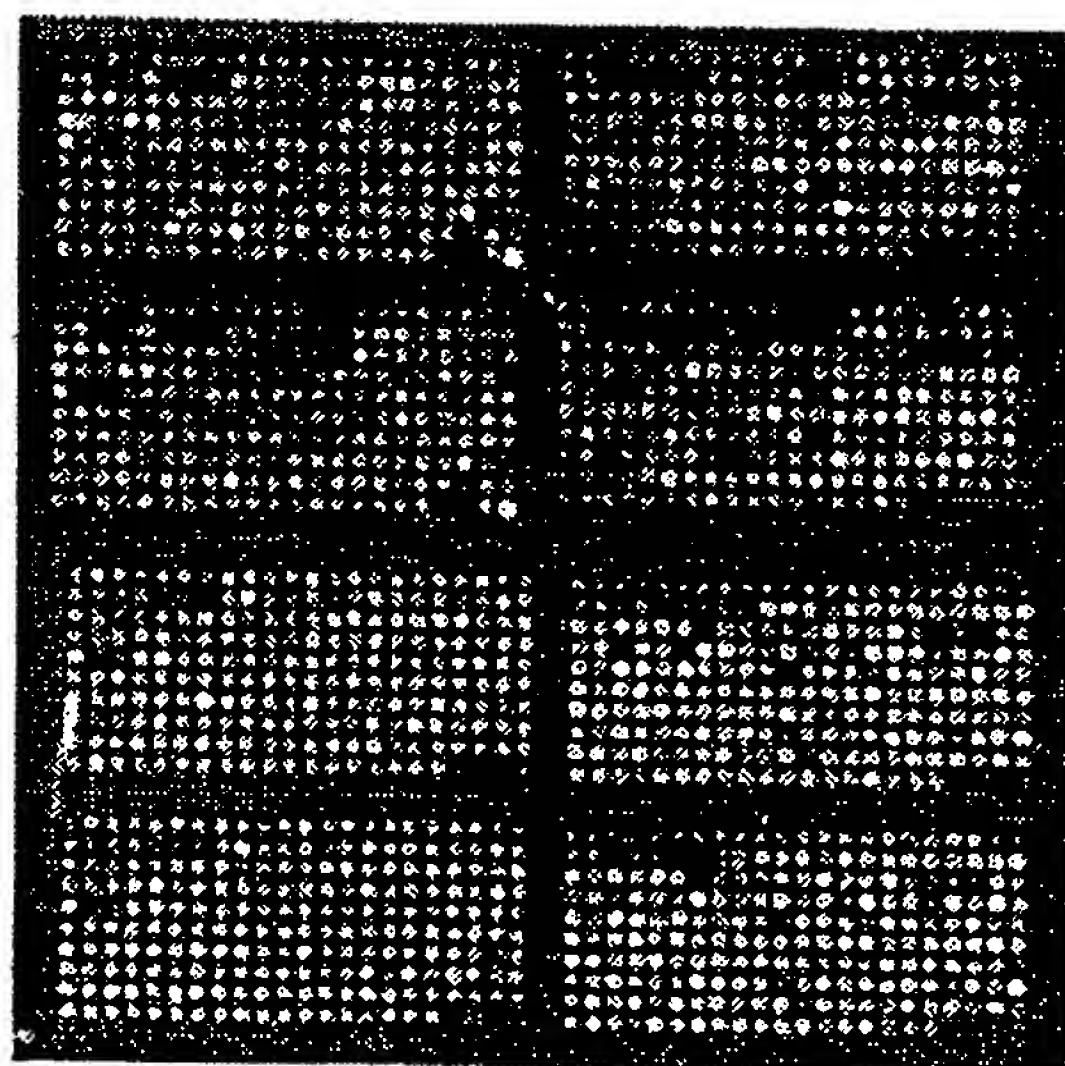


Fig. 6

4/6

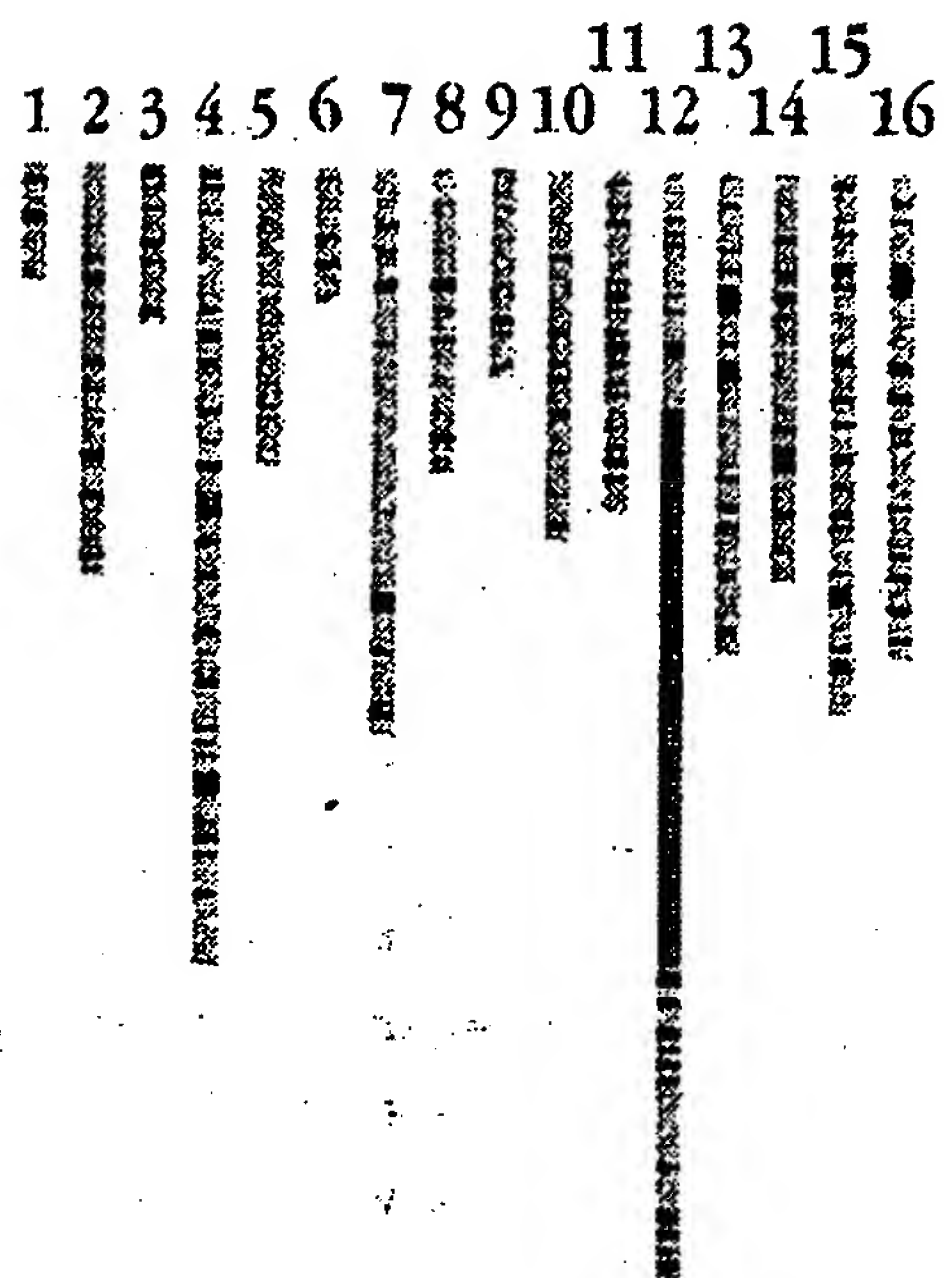


Fig. 7



Fig. 8

5/6

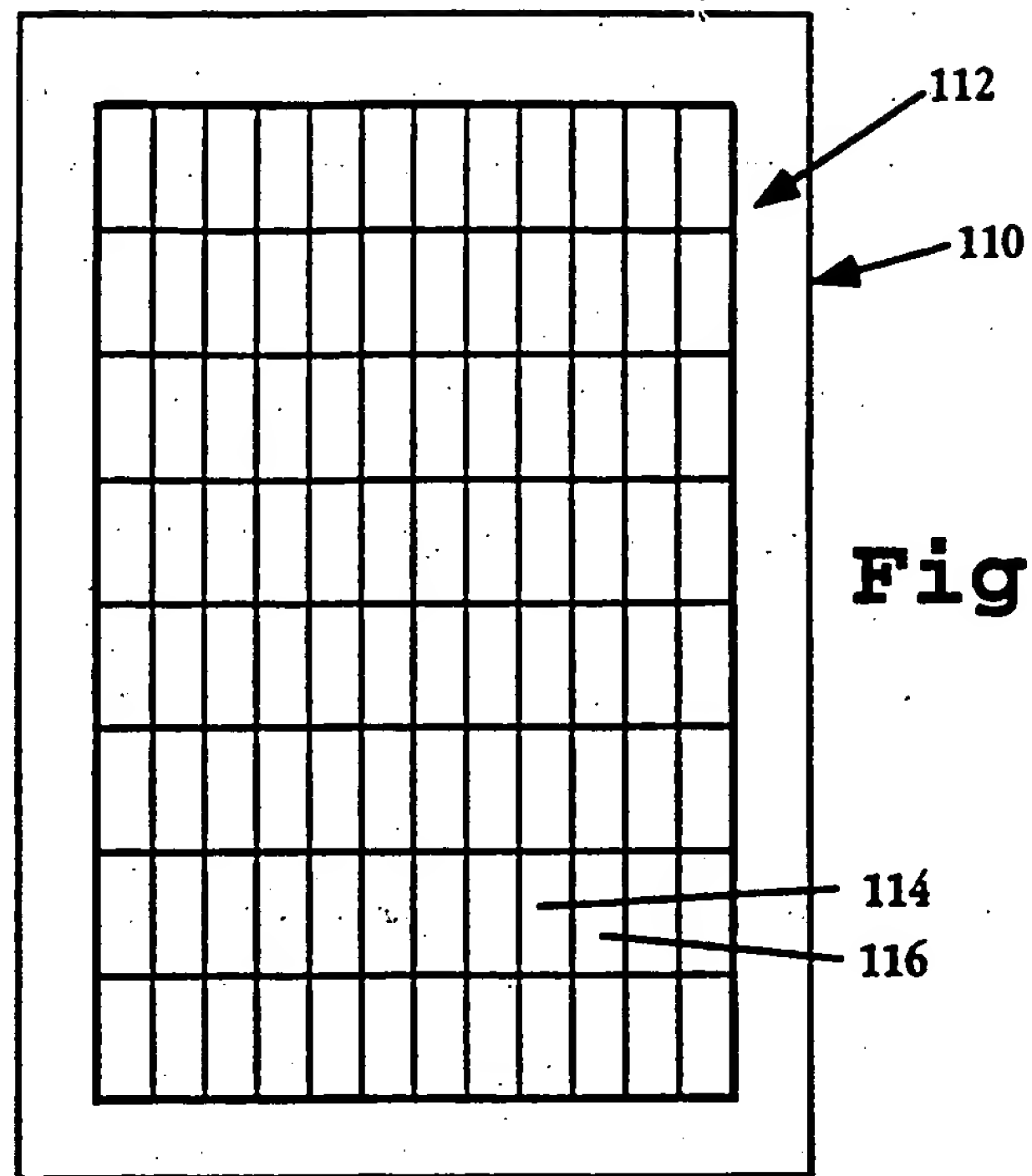


Fig. 9

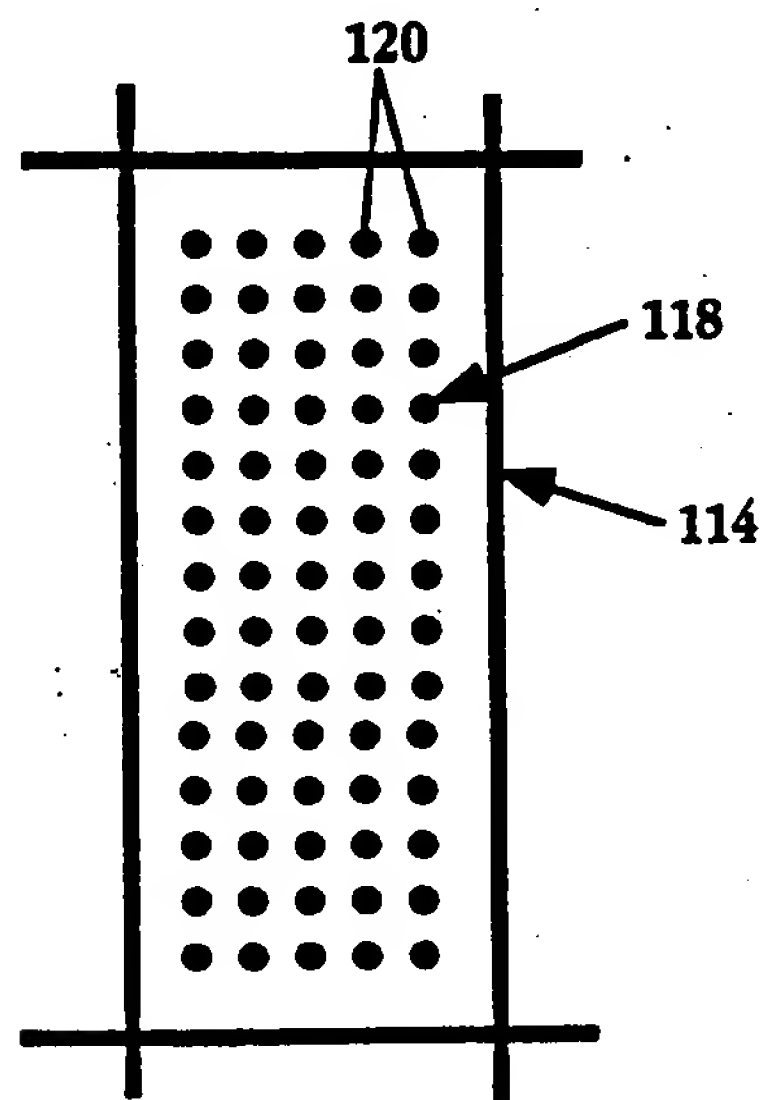


Fig. 10

6/6

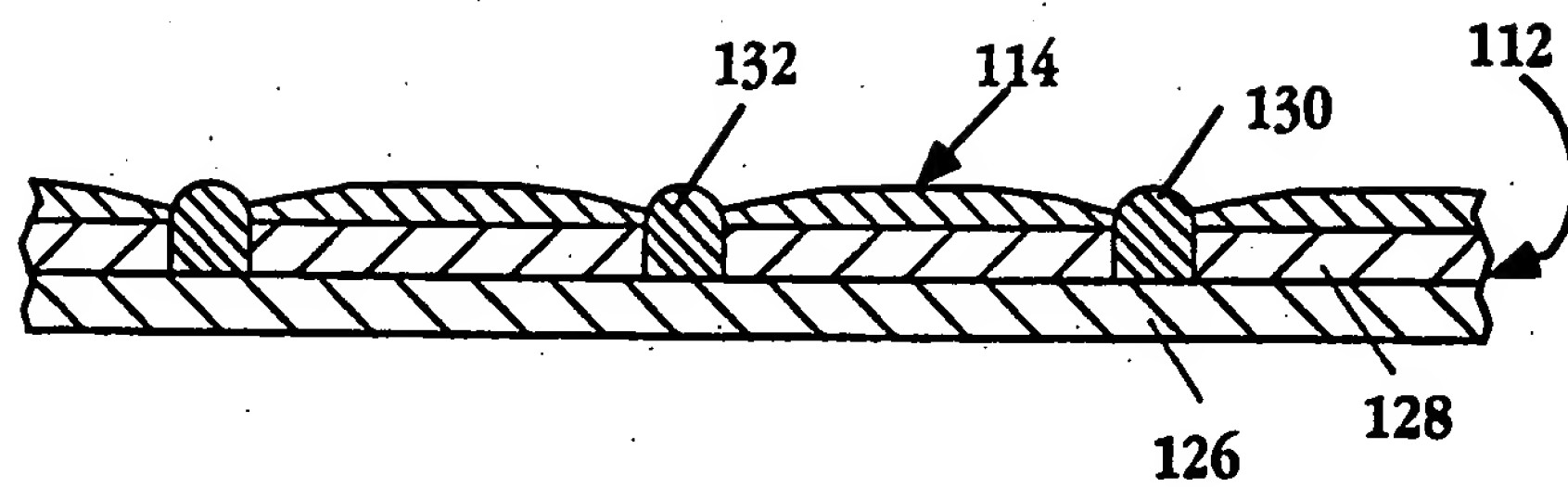


Fig. 11

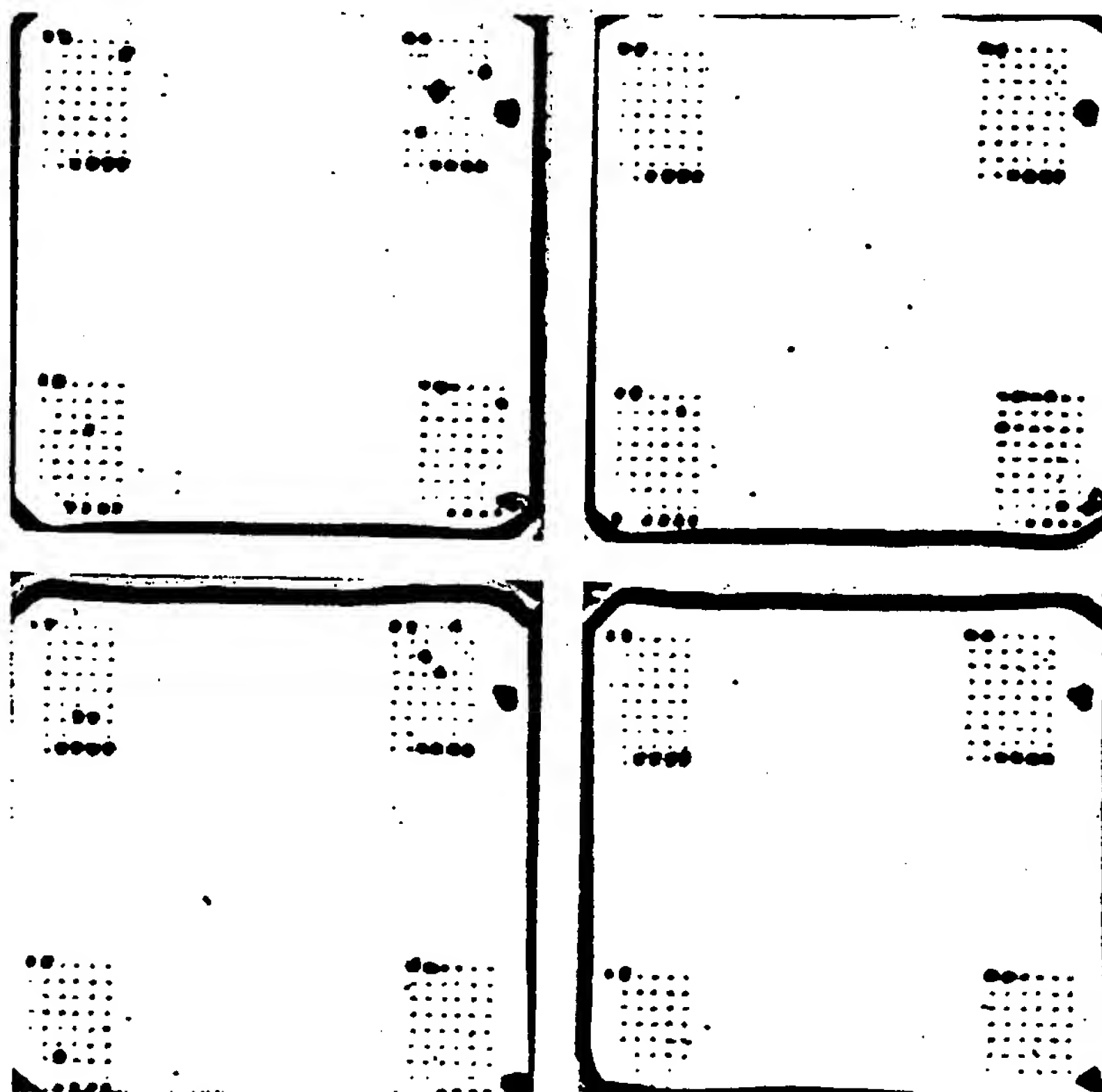


Fig. 12

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US95/07659

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) : G01N 33/543, 33/68

US CL : 435/6; 436/518

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 422/57; 435/4.6.973; 436/518,524,527,531,805,809

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A,P	US, A, 5,338,688 (DEEG ET AL) 16 August 1994, see entire document	1-17
A	US, A, 5,204,268 (MATSUMOTO) 20 April 1993, see entire document.	6-11
A	US, A, 4,071,315 (CHATEAU) 31 January 1978, see entire document.	12-17
A	US, A, 5,100,777 (CHANG) 31 March 1992, see entire document.	12-17
A	US, A, 5,200,312 (OPRANDY) 06 April 1993, see entire document.	12-17

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:	T	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
A document defining the general state of the art which is not considered to be of particular relevance	X	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
E earlier document published on or after the international filing date	Y	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
L document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	&	document member of the same patent family
O document referring to an oral disclosure, use, exhibition or other means		
P document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search

15 SEPTEMBER 1995

Date of mailing of the international search report

06 OCT 1995

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer

CHRISTOPHER CHIN

Telephone No. (703) 308-0196

Discovery and analysis of inflammatory disease-related genes using cDNA microarrays

(inflammation/human genome analysis/gene discovery)

RENU A. HELLER*[†], MARK SCHENA*, ANDREW CHAI*, DARI SHALON[‡], TOD BEDILION[‡], JAMES GILMORE[‡], DAVID E. WOOLLEY[§], AND RONALD W. DAVIS*

*Department of Biochemistry, Beckman Center, Stanford University Medical Center, Stanford, CA 94305; [‡]Synteni, Palo Alto, CA 94306; and [§]Department of Medicine, Manchester Royal Infirmary, Manchester, United Kingdom

Contributed by Ronald W. Davis, December 27, 1996

ABSTRACT cDNA microarray technology is used to profile complex diseases and discover novel disease-related genes. In inflammatory disease such as rheumatoid arthritis, expression patterns of diverse cell types contribute to the pathology. We have monitored gene expression in this disease state with a microarray of selected human genes of probable significance in inflammation as well as with genes expressed in peripheral human blood cells. Messenger RNA from cultured macrophages, chondrocyte cell lines, primary chondrocytes, and synoviocytes provided expression profiles for the selected cytokines, chemokines, DNA binding proteins, and matrix-degrading metalloproteinases. Comparisons between tissue samples of rheumatoid arthritis and inflammatory bowel disease verified the involvement of many genes and revealed novel participation of the cytokine interleukin 3, chemokine Gro α and the metalloproteinase matrix metallo-elastase in both diseases. From the peripheral blood library, tissue inhibitor of metalloproteinase 1, ferritin light chain, and manganese superoxide dismutase genes were identified as expressed differentially in rheumatoid arthritis compared with inflammatory bowel disease. These results successfully demonstrate the use of the cDNA microarray system as a general approach for dissecting human diseases.

The recently described cDNA microarray or DNA-chip technology allows expression monitoring of hundreds and thousands of genes simultaneously and provides a format for identifying genes as well as changes in their activity (1, 2). Using this technology, two-color fluorescence patterns of differential gene expression in the root versus the shoot tissue of *Arabidopsis* were obtained in a specific array of 48 genes (1). In another study using a 1000 gene array from a human peripheral blood library, novel genes expressed by T cells were identified upon heat shock and protein kinase C activation (3).

The technology uses cDNA sequences or cDNA inserts of a library for PCR amplification that are arrayed on a glass slide with high speed robotics at a density of 1000 cDNA sequences per cm². These microarrays serve as gene targets for hybridization to cDNA probes prepared from RNA samples of cells or tissues. A two-color fluorescence labeling technique is used in the preparation of the cDNA probes such that a simultaneous hybridization but separate detection of signals provides the comparative analysis and the relative abundance of specific genes expressed (1, 2). Microarrays can be constructed from specific cDNA clones of interest, a cDNA library, or a select number of open reading frames from a genome sequencing database to allow a large-scale functional analysis of expressed sequences.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Copyright © 1997 by THE NATIONAL ACADEMY OF SCIENCES OF THE USA
0027-8424/97/942150-6\$2.00/0
PNAS is available online at <http://www.pnas.org>.

Because of the wide spectrum of genes and endogenous mediators involved, the microarray technology is well suited for analyzing chronic diseases. In rheumatoid arthritis (RA), inflammation of the joint is caused by the gene products of many different cell types present in the synovium and cartilage tissues plus those infiltrating from the circulating blood. The autoimmune and inflammatory nature of the disease is a cumulative result of genetic susceptibility factors and multiple responses, paracrine and autocrine in nature, from macrophages, T cells, plasma cells, neutrophils, synovial fibroblasts, chondrocytes, etc. Growth factors, inflammatory cytokines (4), and the chemokines (5) are the important mediators of this inflammatory process. The ensuing destruction of the cartilage and bone by the invading synovial tissue includes the actions of prostaglandins and leukotrienes (6), and the matrix degrading metalloproteinases (MMPs). The MMPs are an important class of Zn-dependent metallo-endoproteases that can collectively degrade the proteoglycan and collagen components of the connective tissue matrix (7).

This paper presents a study in which the involvement of select classes of molecules in RA was examined. Also investigated were 1000 human genes randomly selected from a peripheral human blood cell library. Their differential and quantitative expression analysis in cells of the joint tissue, in diseased RA tissue and in inflammatory bowel disease (IBD) tissues was conducted to demonstrate the utility of the microarray method to analyze complex diseases by their pattern of gene expression. Such a survey provides insight not only into the underlying cause of the pathology, but also provides the opportunity to selectively target genes for disease intervention by appropriate drug development and gene therapies.

METHODS

Microarray Design, Development, and Preparation. Two approaches for the fabrication of cDNA microarrays were used in this study. In the first approach, known human genes of probable significance in RA were identified. Regions of the clones, preferably 1 kb in length, were selected by their proximity to the 3' end of the cDNA and for areas of least identity to related and repetitive sequences. Primers were synthesized to amplify the target regions by standard PCR protocols (3). Products were

Abbreviations: RA, rheumatoid arthritis; MMP, matrix-degrading metalloproteinase; IBD, inflammatory bowel disease; LPS, lipopolysaccharide; PMA, phorbol 12-myristate 13-acetate; TNF- α , tumor necrosis factor α ; IL, interleukin; TGF- β , transforming growth factor β ; GCSF, granulocyte colony-stimulating factor; MIP, macrophage inflammatory protein; MIF, migration inhibitory factor; HME, human matrix metallo-elastase; RANTES, regulated upon activation, normal T cell expressed and secreted; Gel, gelatinase; VCAM, vascular cell adhesion molecule; ICE, IL-1 converting enzyme; PUMP, putative metalloproteinase; MnSOD, manganese superoxide dismutase; TIMP, tissue inhibitor of metalloproteinase; MCP, macrophage chemotactic protein.

[†]To whom reprint requests should be sent at the present address: Roche Bioscience, S3-1, 3401 Hillview Avenue, Palo Alto, CA 94304.

verified by gel electrophoresis and purified with Qiaquick 96-well purification kit (Qiagen, Chatsworth, CA), lyophilized (Savant), and resuspended in 5 μ l of 3 \times standard saline citrate (SSC) buffer for arraying. In the second approach, the microarray containing the 1056 human genes from the peripheral blood lymphocyte library was prepared as described (3).

Tissue Specimens. Rheumatoid synovial tissue was obtained from patients with late stage classic RA undergoing remedial synovectomy or arthroplasty of the knee. Synovial tissue was separated from any associated connective tissue or fat. One gram of each synovial specimen was subjected to RNA extraction within 40 min of surgical excision, or explants were cultured in serum-free medium to examine any changes under *in vitro* conditions. For IBD, specimens of macroscopically inflamed lower intestinal mucosa were obtained from patients with Crohn disease undergoing remedial surgery. The hypertrophied mucosal tissue was separated from underlying connective tissue and extracted for RNA.

Cultured Cells. The Mono Mac-6 (MM6) monocytic cells (8) were grown in RPMI medium. Human chondrosarcoma SW1353 cells, primary human chondrocytes, and synoviocytes (9, 10) were cultured in DMEM; all culture media were supplemented with 10% fetal bovine serum, 100 μ g/ml streptomycin, and 500 units/ml penicillin. Treatment of cells with lipopolysaccharide (LPS) endotoxin at 30 ng/ml, phorbol 12-myristate 13-acetate (PMA) at 50 ng/ml, tumor necrosis factor α (TNF- α) at 50 ng/ml, interleukin (IL)-1 β at 30 ng/ml, or transforming growth factor- β (TGF- β) at 100 ng/ml is described in the figure legends.

Fluorescent Probe, Hybridization, and Scanning. Isolation of mRNA, probe preparation, and quantitation with *Arabidopsis* control mRNAs was essentially as described (3) except for the following minor modification. Following the reverse transcriptase step, the appropriate Cy3- and Cy5-labeled samples were pooled; mRNA degraded by heating the sample to 65°C for 10 min with the addition of 5 μ l of 0.5M NaOH plus 0.5 ml of 10 mM EDTA. The pooled cDNA was purified from unincorporated nucleotides by gel filtration in Centri-spin columns (Princeton Separations, Adelphia, NJ). Samples were lyophilized and dissolved in 6 μ l of hybridization buffer (5 \times SSC plus 0.2% SDS). Hybridizations, washes, scanning, quantitation procedures, and pseudocolor representations of fluorescent images have been described (3). Scans for the two fluorescent probes were normalized either to the fluorescence intensity of *Arabidopsis* mRNAs spiked into the labeling reactions (see Figs. 2–4) or to the signal intensity of β -actin and glyceraldehyde-3-phosphate dehydrogenase (GAPDH; see Fig. 5).

RESULTS

Ninety-Six-Genes Microarray Design. The actions of cytokines, growth factors, chemokines, transcription factors, MMPs, prostaglandins, and leukotrienes are well recognized in inflammatory disease, particularly RA (11–14). Fig. 1 displays the selected genes for this study and also includes control cDNAs of housekeeping genes such as β -actin and GAPDH and genes from *Arabidopsis* for signal normalization and quantitation (row A, columns 1–12).

Defining Microarray Assay Conditions. Different lengths and concentrations of target DNA were tested by arraying PCR-

	1	2	3	4	5	6	7	8	9	10	11	12
A	BLANK	BLANK	HAT1 HAT1	HAT1 HAT1	HAT4 HAT4	HAT4 HAT4	HAT22 HAT22	HAT22 HAT22	YES23 YES23	YES23 YES23	BACTIN β -actin	G3PDH G3PDH
B	IL1A IL-1 α	IL1B IL-1 β	IL1RA IL-1RA	IL2 IL-2	IL3 IL-3	IL4 IL-4	IL6 IL-6	IL6R IL-6R	IL7 IL-7	CFOS c-fos	CJUN c-jun	RFRA1 Rat Fra-1
C	IL8 IL-8	IL9 IL-9	IL10 IL-10	ICE ICE	IFNG IFN γ	GCSF G-CSF	MCSF M-CSF	GMCSF GM-CSF	TNFB.1 TNF β	CREL c-rel	NFKB50 NF κ Bp50	NFKB65.1 NF κ Bp65
D	TNFA.1 TNF α	TNFA.2 TNF α	TNFA.3 TNF α	TNFA.4 TNF α	TNFA.5 TNF α	TNFR1.1 TNFR1	TNFR1.2 TNFR1	TNFR1.1 TNFR1	TNFR1.2 TNFR1	NFKB65.2 NF κ Bp65	IKB I κ B	CREB2 CREB2
E	STR1 Strom-1	STR2-3' Strom-2	STR3 Strom-3	COL1 Coll-1	COL1-3' Coll-1.3'	COL2.1 Coll-2	COL2.2 Coll-2	COL3 Coll-3	COX1 Cox-1	COX2 Cox-2	12LO 12-LO	15LO 15-LO
F	GELA.1 Gel-A	GELB Gel-B	HME Elastase	MTMMP MT-MMP	PUMP1 Matrilysin	TIMP1 TIMP-1	TIMP2 TIMP-2	TIMP3 TIMP-3	ICAM1 ICAM-1	VCAM VCAM	5LO.1 5-LO	CPLA2.2 cPLA2
G	EGF EGF	EGFA EGF acidic	EGFB EGF basic	IGFI IGF-I	IGFII IGF-II	TGFA TGF α	TGFB TGF β	PDGFB PDGF β	CALCTN Calctonin	GH1 GH-1	GRO GRO1 α	GGR GR
H	MCP1.1 MCP-1	MCP1.1 MCP-1	MIP1A MIP-1 α	MIP1B MIP-1 β	MIF MIF	RANTES RANTES	INOS INOS	LDLR LDLR	ALU.1 IL-10	ALU.2 TNFRp70	ALU.3 IL-10	POLYA LDLR

A. thaliana controls

Human controls

Cytokines and related genes

Transcription factors and related genes

MMP's and related genes

Chemokines

Growth factors and related genes

Other genes

FIG. 1. Ninety-six-element microarray design. The target element name and the corresponding gene are shown in the layout. Some genes have more than one target element to guarantee specificity of signal. For TNF the targets represent decreasing lengths of 1, 0.8, 0.6, 0.4, and 0.2 kb from left to right.

amplified products ranging from 0.2 to 1.2 kb at concentrations of 1 $\mu\text{g}/\mu\text{l}$ or less. No significant difference in the signal levels was observed within this range of target size and only with 0.2-kb length was a signal reduced upon an 8-fold dilution of the 1 $\mu\text{g}/\mu\text{l}$ sample (data not shown). In this study the average length of the targets was 1 kb, with a few exceptions in the range of ≈ 300 bp, arrayed at a concentration of 1 $\mu\text{g}/\mu\text{l}$. Normally one PCR provided sufficient material to fabricate up to 1000 microarray targets.

In considering positional effects in the development of the targets for the microarrays, selection was biased toward the 3' proximal regions, because the signal was reduced if the target fragment was biased toward the 5' end (data not shown). This result was anticipated since the hybridizing probe is prepared by reverse transcription with oligo(dT)-primed mRNA and is richer in 3' proximal sequences. Cross-hybridizations of probes to targets of a gene family were analyzed with the matrix metal-

loproteinases as the example because they can show regions of sequence identities of greater than 70%. With collagenase-1 (Col-1) and collagenase-2 (Col-2) genes as targets with up to 70% sequence identity, and stromelysin-1 (Strom-1) and stromelysin-2 (Strom-2) genes with different degrees of identity, our results showed that a short region of overlap, even with 70–90% sequence identity, produced a low level of cross-hybridization. However, shorter regions of identity spread over the length of the target resulted in cross-hybridization (data not shown). For closely related genes, targets were designed by avoiding long stretches of homology. For members of a gene family two or more target regions were included to discriminate between specificity of signal versus cross-hybridization.

Monitoring Differential Expression in Cultured Cell Lines. In RA tissue, the monocyte/macrophage population plays a prominent role in phagocytic and immunomodulatory activities. Typ-

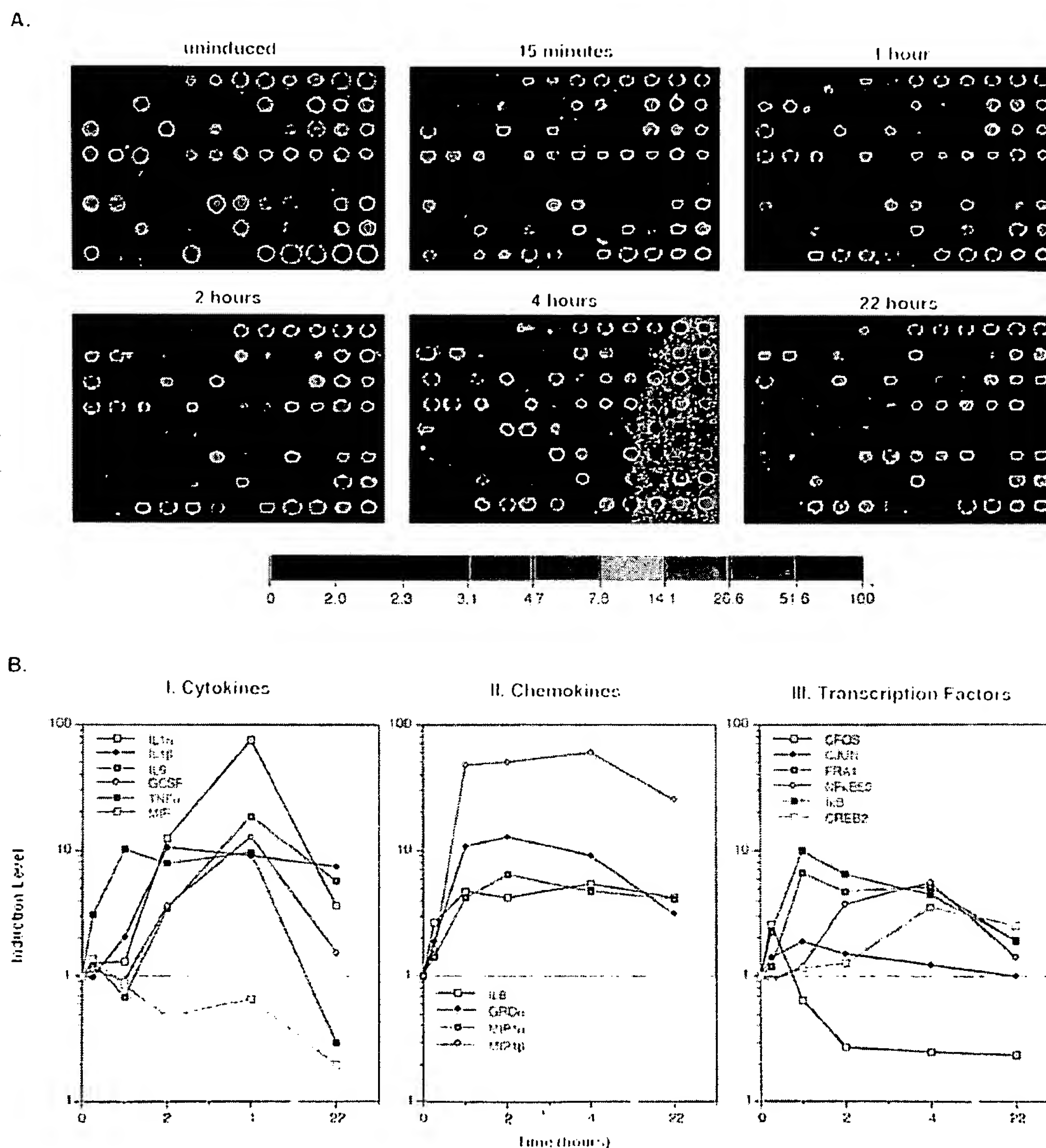


FIG. 2. Time course for LPS/PMA-induced MM6 cells. Array elements are described in Fig. 1. (A) Pseudocolor representations of fluorescent scans correspond to gene expression levels at each time point. The array is made up of 8 *Arabidopsis* control targets and 86 human cDNA targets, the majority of which are genes with known or suspected involvement in inflammation. The color bars provide a comparative calibration scale between arrays and are derived from the *Arabidopsis* mRNA samples that are introduced in equal amounts during probe preparation. Fluorescent probes were made by labeling mRNA from untreated MM6 cells or LPS and PMA treated cells. mRNA was isolated at indicated times after induction. (B I–III) The two-color samples were cohybridized, and microarray scans provided the data for the levels of select transcripts at different time points relative to abundance at time zero. The analysis was performed using normalized data collected from 8-bit images.

ically these cells, when triggered by an immunogen, produce the proinflammatory cytokines TNF and IL-1. We have used the monocyte cell line MM6 and monitored changes in gene expression upon activation with LPS endotoxin, a component of Gram-negative bacterial membranes, and PMA, which augments the action of LPS on TNF production (15). RNA was isolated at different times after induction and used for cDNA probe preparation. From this time course it was clear that TNF expression was induced within 15 min of treatment, reached maximum levels in 1 hr, remained high until 4 hr and subsequently declined (Fig. 2A). Many other cytokine genes were also transiently activated, such as IL-1 α and - β , IL-6, and granulocyte colony-stimulating factor (GCSF). Prominent chemokines activated were IL-8, macrophage inflammatory protein (MIP)-1 β , more so than MIP-1 α , and Gro α or melanoma growth stimulatory factor. Migration inhibitory factor (MIF) expressed in the uninduced state declined in LPS-activated cells. Of the immediate early genes, the noticeable ones were *c-fos*, *fra-1*, *c-jun*, NF- κ Bp50, and I κ B, with *c-rel* expression observed even in the uninduced state (Fig. 2B). These expression patterns are consistent with reported patterns of activation of certain LPS- and PMA-induced genes (12). Demonstrated here is the unique ability of this system to allow parallel visualization of a large number of gene activities over a period of time.

SW1353 cells is a line derived from malignant tumors of the cartilage and behaves much like the chondrocytes upon stimulation with TNF and IL-1 in the expression of MMPs (9). In addition to confirming our earlier observations with Northern blots on Strom-1, Col-1, and Col-3 expression (9), gelatinase (Gel) A, putative metalloproteinase (PUMP)-1 membrane-

type matrix metalloproteinase, tissue inhibitors of matrix metalloproteinases or tissue inhibitor of metalloproteinase 1 (TIMP-1), -2, and -3 were also expressed by these cells together with the human matrix metallo-elastase (HME; Fig. 3A). HME induction was estimated to be ≈ 50 -fold and was greater than any of the other MMPs examined (Fig. 3B). This result was unexpected because HME is reportedly expressed only by alveolar macrophage and placental cells (16). Expression of the cytokines and chemokines, IL-6, IL-8, MIF, and MIP-1 β was also noted. A variety of other genes, including certain transcription factors, were also up-regulated (Fig. 3), but the overall time-dependent expression of genes in the SW1353 cells was qualitatively distinct from the MM6 cells.

Quantitation of differential gene expression (Figs. 2B and 3B) was achieved with the simultaneous hybridization of Cy3-labeled cDNA from untreated cells and Cy5-labeled cDNA from treated samples. The estimated increases in expression from these microarrays for a select number of genes including IL-1 β , IL-8, MIP-1 β , TNF, HME, Col-1, Col-3, Strom-1, and Strom-2 were compared with data collected from dot blot analysis. Results (not shown) were in close agreement and confirmed our earlier observations on the use of the microarray method for the quantitation of gene expression (3).

Expression Profiles in Primary Chondrocytes and Synoviocytes of Human RA Tissue. Given the sensitivity and the specificity of this method, expression profiles of primary synoviocytes and chondrocytes from diseased tissue were examined. Without prior exposure to inducing agents, low level expression of *c-jun*, GCSF, IL-3, TNF- β , MIF, and RANTES (regulated upon activation, normal T cell expressed and secreted) was seen as well as expression of MMPs, GelA, Strom-1, Col-1, and the three TIMPs. In this case, Col-2 hybridization was considered to be nonspecific because the second Col-2 target taken from the 3' end of the gene gave no

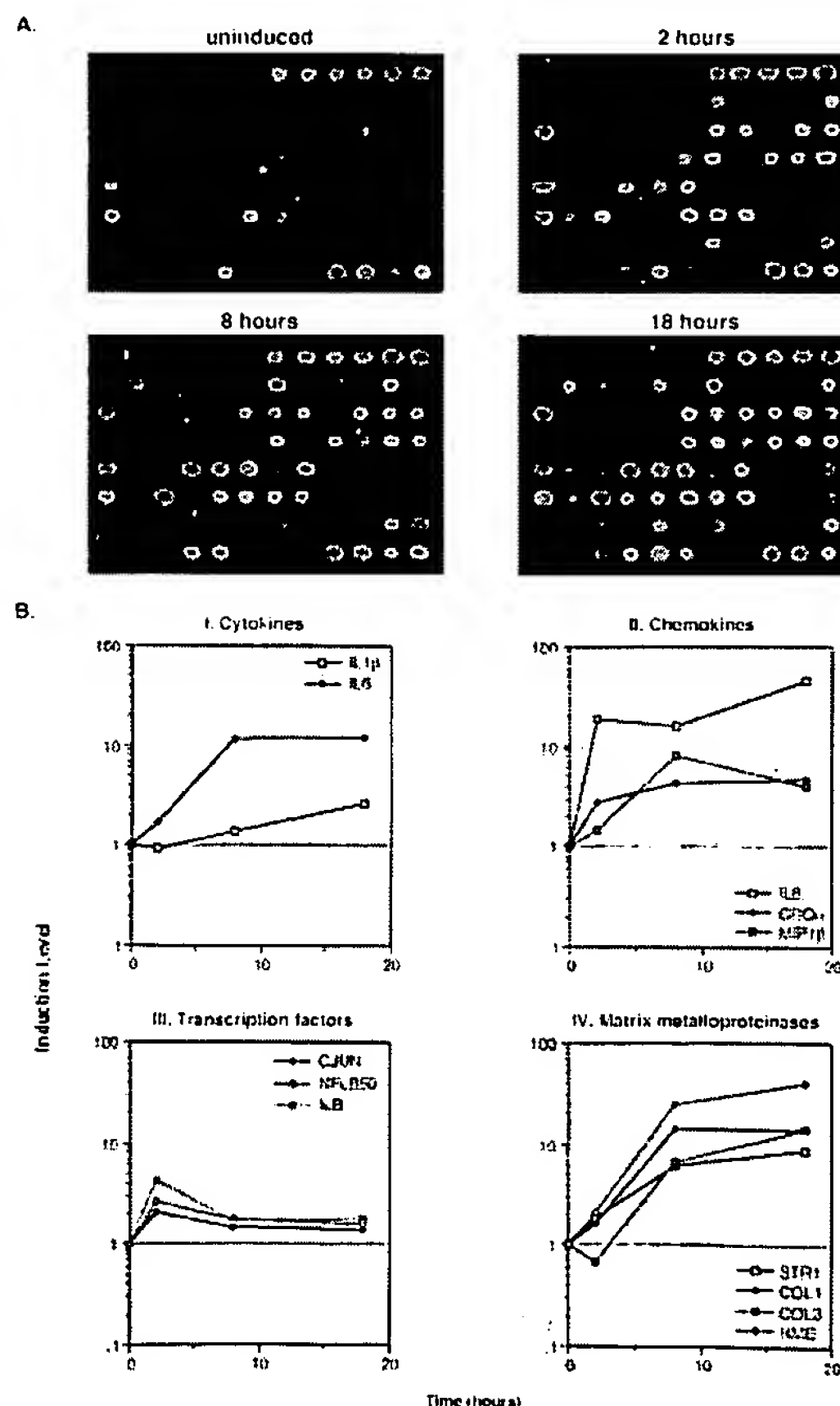


FIG. 3. Time course for IL-1 β and TNF-induced SW1353 cells using the inflammation array (Fig. 1). (A) Pseudocolor representation of fluorescent scans correspond to gene expression levels at each time point. (B I–IV) Relative levels of selected genes at different time points compared with time zero.

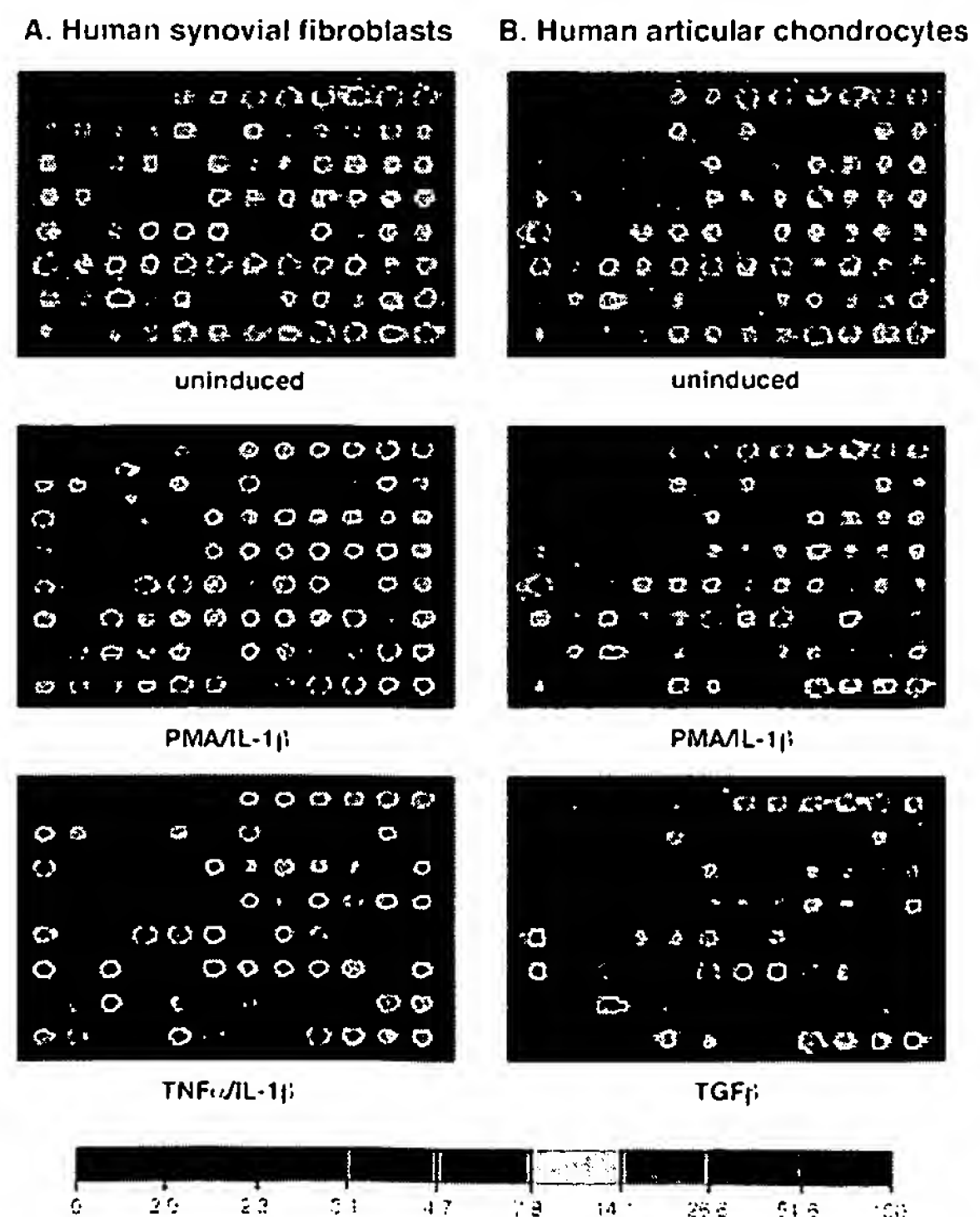


FIG. 4. Expression profiles for early passage primary synoviocytes and chondrocytes isolated from RA tissue, cultured in the presence of 10% fetal calf serum and activated with PMA and IL-1 β , or TNF and IL-1 β , or TGF- β for 18 hr. The color bars provide a comparative calibration scale between arrays and are derived from the *Arabidopsis* mRNA samples that are introduced in equal amounts during probe preparation.

signal. Treatment more so with PMA and IL-1, than TNF and IL-1, produced a dramatic up-regulation in expression of several genes in both of these primary cell types. These genes are as follows: the cytokine IL-6, the chemokines IL-8 and Gro-1 α , and the MMPs; Strom-1, Col-1, Col-3, and HME; and the adhesion molecule, vascular cell adhesion molecule 1 (VCAM-1). The surprise again is HME expression in these primary cells, for reasons discussed above. From these results, the expression profiles of synoviocytes and the chondrocytes appear very similar; the differences are more quantitative than qualitative. Treatment of the primary chondrocytes with the anabolic growth factor TGF- β had an interesting profile in that it produced a remarkable down-regulation of genes expressed in both the untreated and induced state (Fig. 4).

Given the demonstrated effectiveness of this technology, a comparative analysis of two different inflammatory disease states was conducted with probes made from RA tissue and IBD samples. RA samples were from late stage rheumatoid synovial tissue, and IBD specimens were obtained from inflamed lower intestinal mucosa of patients with Crohn disease. With both the 96-element known gene microarray and the 1000-gene microarray of cDNAs selected from a peripheral human blood cell library (3), distinct differences in gene expression patterns were evident. On the 96-gene array, RA tissue samples from different affected individuals gave similar profiles (data not shown) as did different samples from the same individual (Fig. 5). These patterns were notably similar to those observed with primary synoviocytes and chondrocytes (Fig. 4). Included in the list of prominently up-regulated genes are IL-6, the MMPs Strom-1, Col-1, GelA, HME, and in

certain samples PUMP, TIMPs, particularly TIMP-1 and TIMP-3, and the adhesion molecule VCAM. Discernible levels of macrophage chemotactic protein 1 (MCP-1), MIF and RANTES were also noted. IBD samples were in comparison, rather subdued although IL-1 converting enzyme (ICE), TIMP-1, and MIF were notable in all the three different IBD samples examined here. In IBD-A, one of three individual samples, ICE, VCAM, Gro α , and MMP expression was more pronounced than in the others.

We also made use of a peripheral blood cDNA library (3) to identify genes expressed by lymphocytes infiltrating the inflamed tissues from the circulating blood. With the 1046-element array of randomly selected cDNAs from this library, probes made from RA and IBD samples showed hybridizations to a large number of genes. Of these, many were common between the two disease tissues while others were differentially expressed (data not shown). A complete survey of these genes was beyond the scope of this study, but for this report we picked three genes that were up-regulated in the RA tissue relative to IBD. These cDNAs were sequenced and identified by comparison to the GenBank database. They are TIMP-1, apoferritin light chain, and manganese superoxide dismutase (MnSOD). Differential expression of MnSOD was only observed in samples of RA tissue explants maintained in growth medium without serum for anywhere between 2 to 16 hr. These results also indicate that the expression profile of genes can be altered when explants are transferred to culture conditions.

DISCUSSION

The speed, ease, and feasibility of simultaneously monitoring differential expression of hundreds of genes with the cDNA microarray based system (1–3) is demonstrated here in the analysis of a complex disease such as RA. Many different cell types in the RA tissue; macrophages, lymphocytes, plasma cells, neutrophils, synoviocytes, chondrocytes, etc. are known to contribute to the development of the disease with the expression of gene products known to be proinflammatory. They include the cytokines, chemokines, growth factors, MMPs, eicosanoids, and others (7, 11–14), and the design of the 96-element known gene microarray was based on this knowledge and depended on the availability of the genes. The technology was validated by confirming earlier observations on the expression of TNF by the monocyte cell line MM6, and of Col-1 and Col-3 expression in the chondrosarcoma cells and articular chondrocytes (9, 12). In our time-dependent survey the chronological order of gene activities in and between gene families was compared and the results have provided unprecedented profiles of the cytokines (TNF, IL-1, IL-6, GCSF, and MIF), chemokines (MIP-1 α , MIP-1 β , IL-8, and Gro-1), certain transcription factors, and the matrix metalloproteinases (GelA, Strom-1, Col-1, Col-3, HME) in the macrophage cell line MM6 and in the SW1353 chondrosarcoma cells.

Earlier reports of cytokine production in the diseased state had established a model in which TNF is a major participant in RA. Its expression reportedly preceded that of the other cytokines and effector molecules (4). Our results strongly support these results as demonstrated in the time course of the MM6 cells where TNF induction preceded that of IL-1 α and IL- β followed by IL-6 and GCSF. These expression profiles demonstrate the utility of the microarrays in determining the hierarchy of signaling events.

In the SW1353 chondrosarcoma cells, all the known MMPs and TIMPs were examined simultaneously. HME expression was discovered, which previously had been observed in only the stromal cells and alveolar macrophages of smoker's lungs and in placental tissue. Its presence in cells of the RA tissue is meaningful because its activity can cause significant destruction of elastin and basement membrane components (16, 17). Expression profiles of synovial fibroblasts and articular chondrocytes were remarkably similar and not too different from the SW1353 cells, indicating that the fibroblast and the chondrocyte can play equally aggressive roles in joint erosion. Prominent genes expressed were

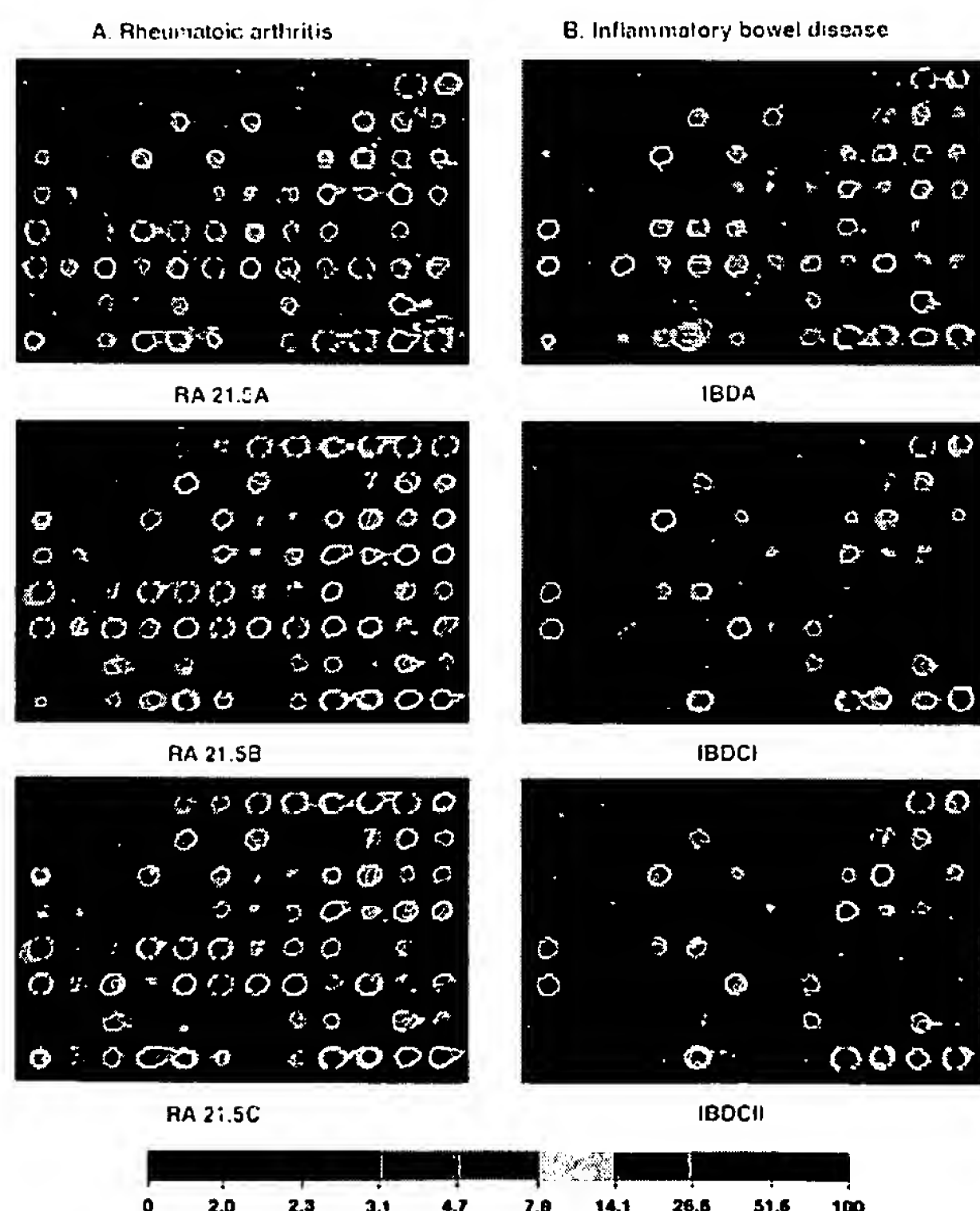


FIG. 5. Expression profiles of RA tissue (A) and IBD tissue (B). mRNA from RA tissue samples obtained from the same individual was isolated directly after excision (RA 21.5A) or maintained in culture without serum for 2 hr (RA 21.5B) or for 6 hr (RA 21.5C). Profiles from tissue samples of two other individuals (data not shown) were remarkably similar to the ones shown here. IBD-A and IBD-CI are from mRNA samples prepared directly after surgery from two separate individuals. For the IBD-CII probe, the tissue sample was cultured in medium without serum for 2 hr before mRNA preparation.

the MMPs, but chemokines and cytokines were also produced by these cells. The effect of the anabolic growth factor TGF- β was profoundly evident in demonstrating the down regulation of these catabolic activities.

RA tissue samples undeniably reflected profiles similar to the cell types examined. Active genes observed were IL-3, IL-6, ICE, the MMPs including HME and TIMPs, chemokines IL-8, Gro α , MIP, MIF, and RANTES, and the adhesion molecule VCAM. Of the growth factors, fibroblast growth factor β was observed most frequently. In comparison, the expression patterns in the other inflammatory state (i.e., IBD) were not as marked as in the RA samples, at least as obtained from the tissue samples selected for this study.

As an alternative approach, the 1046 cDNA microarray of randomly selected genes from a lymphocyte library was used to identify genes expressed in RA tissue (3). Many genes on this array hybridized with probes made from both RA and IBD tissue samples. The results are not surprising because inflammatory tissue is abundantly supplied with cell types infiltrating from the circulating blood, made apparent also by the high levels of chemokine expression in RA tissue. Because of the magnitude of the effort required to identify all the hybridized genes, we have for this report chosen to describe only three differentially expressed genes mainly to verify this method of analysis.

Of the large number of genes observed here, a fair number were already known as active participants in inflammatory disease. These are TNF, IL-1, IL-6, IL-8, GCSF, RANTES, and VCAM. The novel participants not previously reported are HME, IL-3, ICE, and Gro α . With our discovery of HME expression in RA, this gene becomes a target for drug intervention. ICE is a cysteine protease well known for its IL-1 β processing activity (18), and recognized for its role in apoptotic cell death (19). Its expression in RA tissue is intriguing. IL-3 is recognized for its growth-promoting activity in hematopoietic cell lineages, is a product of activated T cells (20), and its expression in synovio-cytes and chondrocytes of RA tissue is a novel observation.

Like IL-8, Gro α , is a C-X-C subgroup chemokine and is a potent neutrophil and basophil chemoattractant. It down-regulates the expression of types I and III interstitial collagens (21, 22) and is seen here produced by the MM6 cells, in primary synovio-cytes, and in RA tissue. With the presence of RANTES, MCP, and MIP-1 β , the C-C chemokines (23) migration and infiltration of monocytes, particularly T cells, into the tissue is also enhanced (5) and aid in the trafficking and recruitment of leukocytes into the RA tissue. Their activation, phagocytosis, degranulation, and respiratory bursts could be responsible for the induction of MnSOD in RA. MnSOD is also induced by TNF and IL-1 and serves a protective function against oxidative damage. The induction of the ferritin light chain encoding gene in this tissue may be for reasons similar to those for MnSOD. Ferritin is the major intracellular iron storage protein and it is responsive to intracellular oxidative stress and reactive oxygen intermediates generated during inflammation (24, 25). The active expression of TIMP-1 in RA tissue, as detected by the 1000-element array, is no surprise because our results have repeatedly shown TIMP-1 to be expressed in the constitutive and induced states of RA cells and tissues.

The suitability of the cDNA microarray technology for profiling diseases and for identifying disease related genes is well documented here. This technology could provide new

targets for drug development and disease therapies, and in doing so allow for improved treatment of chronic diseases that are challenging because of their complexity.

We would like to thank the following individuals for their help in obtaining reagents or providing cDNA clones to use as templates in target preparation: N. Arai, P. Cannon, D. R. Cohen, T. Curran, V. Dixit, D. A. Geller, G. I. Goldberg, M. Karin, M. Lotz, L. Matrisian, G. Nolan, C. Lopez-Otin, T. Schall, S. Shapiro, I. Verma, and H. Van Wart. Support for R.W.D., M.S., and R.A.H. was provided by the National Institutes of Health (Grants R37HG00198 and HG00205).

1. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995) *Science* **270**, 467–470.
2. Shalon, D., Smith, S. & Brown, P. O. (1996) *Genome Res.* **6**, 639–645.
3. Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O. & Davis, R. W. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 10614–10619.
4. Feldmann, M., Brennan F. M. & Maini, R. N. (1996) *Rheumatoid Arthritis Cell* **85**, 307–310.
5. Schall, T. J. (1994) in *The Cytokine Handbook*, ed. Thomson, A. W. (Academic, New York), 2nd Ed., pp. 410–460.
6. Lotz, M. F., Blanco, J., Von Kempis, J., Dudler, J., Maier, R., Villiger P. M. & Geng, Y. (1995) *J. Rheumatol.* **22**, Supplement 43, 104–108.
7. Birkedal-Hansen, H., Moore, W. G. I., Bodden, M. K., Windsor, L. J., Birkedal-Hansen, B., DeCarlo, A. & Engler, J. A. (1993) *Crit. Rev. Oral Biol. Med.* **4**, 197–250.
8. Zeigler-Heitbrock, H. W. L., Thiel, E., Futterer, A., Volker, H., Wirtz, A. & Reithmuller, G. (1988) *Int. J. Cancer* **41**, 456–461.
9. Borden, P., Solymar, D., Sucharczuk, A., Lindman, B., Cannon, P. & Heller, R. A. (1996) *J. Biol. Chem.* **271**, 23577–23581.
10. Gadher, S. J. & Woolley, D. E. (1987) *Rheumatol. Int.* **7**, 13–22.
11. Harris, E. D., Jr. (1990) *New Engl. J. Med.* **322**, 1277–1289.
12. Firestein, G. S. (1996) in *Textbook of Rheumatology*, eds. Kelly, W. N., Harris, E. D., Ruddy, S. & Sledge, C. B. (Saunders, Philadelphia), 5th Ed. pp. 5001–5047.
13. Alvaro-Garcia, J. M., Zvaifler, Nathan J., Brown, C. B., Kaushansky, K. & Firestein, Gary S. (1991) *J. Immunol.* **146**, 3365–3371.
14. Firestein, G. S., Alvaro-Garcia, J. M. & Maki, R. (1990) *J. Immunol.* **144**, 3347–3352.
15. Pradines-Figueres, A. & Raetz, C. R. H. (1992) *J. Biol. Chem.* **267**, 23261–23268.
16. Shapiro, S. D., Kobayashi, D. L. & Ley, T. J. (1993) *J. Biol. Chem.* **268**, 23824–23829.
17. Shipley, M. J., Wesselschmidt, R. L., Kobayashi, D. K., Ley, T. J. & Shapiro, S. D. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 3042–3046.
18. Cerreti, D. P., Kozlosky, C. J., Mosley, B., Nelson, N., Van Ness, K., Greenstreet, T. A., March, C. J., Kronheim, S. R., Druck, T., Cannizaro, L. A., Huebner, K. & Black, R. A. (1992) *Science* **256**, 97–100.
19. Miura, M., Zhu, H., Rotello, R., Hartweig, E. A. & Yuan, J. (1993) *Cell* **75**, 653–660.
20. Arai, K., Lee, F., Miyajima, A., Shoichiro, M., Arai, N. & Takashi, Y. (1990) *Annu. Rev. Biochem.* **59**, 783–836.
21. Geiser, T., Dewald, B., Ehrenguber, M. U., Lewis, I. C. & Baggiolini, M. (1993) *J. Biol. Chem.* **268**, 15419–15424.
22. Unemori, E. N., Amento, E. P., Bauer, E. A. & Horuk, R. (1993) *J. Biol. Chem.* **268**, 1338–1342.
23. Robinson, E., Keystone, E. C., Schall, T. J., Gillet, N. & Fish, E. N. (1995) *Clin. Exp. Immunol.* **101**, 398–407.
24. Roeser, H. (1980) in *Iron Metabolism in Biochemistry and Medicine*, eds. Jacobs, A. & Worwood, M. (Academic, New York), Vol. 2, pp. 605–640.
25. Kwak, E. L., Larochelle, D. A., Beaumont, C., Torti, S. V. & Torti, F. M. (1995) *J. Biol. Chem.* **270**, 15285–15293.



PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁶ : C12Q 1/68, C07H 21/04	A1	(11) International Publication Number: WO 97/13877 (43) International Publication Date: 17 April 1997 (17.04.97)
(21) International Application Number: PCT/US96/16342 (22) International Filing Date: 11 October 1996 (11.10.96) (30) Priority Data: PCT/US95/12791 12 October 1995 (12.10.95) WO (34) Countries for which the regional or international application was filed: US et al. PCT/US96/09513 6 June 1996 (06.06.96) WO (34) Countries for which the regional or international application was filed: US et al. (60) Parent Application or Grant (63) Related by Continuation US Not furnished (CIP) Filed on Not furnished (71) Applicant (for all designated States except US): LYNX THERAPEUTICS, INC. [US/US]; 3832 Bay Center Place, Hayward, CA 94545 (US). (72) Inventor; and (75) Inventor/Applicant (for US only): MARTIN, David, W. [US/US]; Lynx Therapeutics, Inc., 3832 Bay Center Place, Hayward, CA 94545 (US).		(74) Agent: POWERS, Vincent, M.; Dehlinger & Associates, Post Office Box 60850, Palo Alto, CA 94306-0850 (US). (81) Designated States: AU, CA, CZ, EE, FI, HU, JP, KR, LT, LV, NO, NZ, PL, RU, SG, US, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE). Published <i>With international search report.</i> <i>Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.</i>
(54) Title: MEASUREMENT OF GENE EXPRESSION PROFILES IN TOXICITY DETERMINATION (57) Abstract A method is provided for assessing the toxicity of a compound in a test organism by measuring gene expression profiles of selected tissues. Gene expression profiles are measured by massively parallel signature sequencing of cDNA libraries constructed from mRNA extracted from the selected tissues. Gene expression profiles provide extensive information on the effects of administering a compound to a test organism in both acute toxicity tests and in prolonged and chronic toxicity tests.		

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AM	Armenia	GB	United Kingdom	MW	Malawi
AT	Austria	GE	Georgia	MX	Mexico
AU	Australia	GN	Guinea	NE	Niger
BB	Barbados	GR	Greece	NL	Netherlands
BE	Belgium	HU	Hungary	NO	Norway
BF	Burkina Faso	IE	Ireland	NZ	New Zealand
BG	Bulgaria	IT	Italy	PL	Poland
BJ	Benin	JP	Japan	PT	Portugal
BR	Brazil	KE	Kenya	RO	Romania
BY	Belarus	KG	Kyrgyzstan	RU	Russian Federation
CA	Canada	KP	Democratic People's Republic of Korea	SD	Sudan
CF	Central African Republic	KR	Republic of Korea	SE	Sweden
CG	Congo	KZ	Kazakhstan	SG	Singapore
CH	Switzerland	LI	Liechtenstein	SI	Slovenia
CI	Côte d'Ivoire	LK	Sri Lanka	SK	Slovakia
CM	Cameroon	LR	Liberia	SN	Senegal
CN	China	LT	Lithuania	SZ	Swaziland
CS	Czechoslovakia	LU	Luxembourg	TD	Chad
CZ	Czech Republic	LV	Latvia	TG	Togo
DE	Germany	MC	Monaco	TJ	Tajikistan
DK	Denmark	MD	Republic of Moldova	TT	Trinidad and Tobago
EE	Estonia	MG	Madagascar	UA	Ukraine
ES	Spain	ML	Mali	UG	Uganda
FI	Finland	MN	Mongolia	US	United States of America
FR	France	MR	Mauritania	UZ	Uzbekistan
GA	Gabon			VN	Viet Nam

MEASUREMENT OF GENE EXPRESSION PROFILES **IN TOXICITY DETERMINATION**

5

Field of the Invention

The invention relates generally to methods for detecting and monitoring phenotypic changes in in vitro and in vivo systems for assessing and/or determining the toxicity of chemical compounds, and more particularly, the invention relates to a method for detecting and monitoring changes in gene expression patterns in in vitro and in vivo systems for determining the toxicity of drug candidates.

BACKGROUND

The ability to rapidly and conveniently assess the toxicity of new compounds is extremely important. Thousands of new compounds are synthesized every year, and many are introduced to the environment through the development of new commercial products and processes, often with little knowledge of their short term and long term health effects. In the development of new drugs, the cost of assessing the safety and efficacy of candidate compounds is becoming astronomical: It is estimated that the pharmaceutical industry spends an average of about 300 million dollars to bring a new pharmaceutical compound to market, e.g. Biotechnology, 13: 226-228 (1995). A large fraction of these costs are due to the failure of candidate compounds in the later stages of the developmental process. That is, as the assessment of a candidate drug progresses from the identification of a compound as a drug candidate--for example, through relatively inexpensive binding assays or in vitro screening assays, to pharmacokinetic studies, to toxicity studies, to efficacy studies in model systems, to preliminary clinical studies, and so on, the costs of the associated tests and analyses increases tremendously. Consequently, it may cost several tens of millions of dollars to determine that a once promising candidate compound possesses a side effect or cross reactivity that renders it commercially infeasible to develop further. A great challenge of pharmaceutical development is to remove from further consideration as early as possible those compounds that are likely to fail in the later stages of drug testing.

Drug development programs are clearly structured with this objective in mind; however, rapidly escalating costs have created a need to develop even more stringent and less expensive screens in the early stages to identify false leads as soon as possible. Toxicity assessment is an area where such improvements may be made, for both drug development and for assessing the environmental, health, and safety effects of new compounds in general.

Typically the toxicity of a compound is determined by administering the compound to one or more species of test animal under controlled conditions and by monitoring the effects on a wide range of parameters. The parameters include such things as blood chemistry, weight gain or loss, a variety of behavioral patterns, muscle tone, body temperature, respiration rate, lethality, and the like, which collectively provide a measure of the state of health of the test animal. The degree of deviation of such parameters from their normal ranges gives a measure of the toxicity of a compound. Such tests may be designed to assess the acute, prolonged, or chronic toxicity of a compound. In general, acute tests involve administration of the test chemical on one occasion. The period of observation of the test animals may be as short as a few hours, although it is usually at least 24 hours and in some cases it may be as long as a week or more. In general, prolonged tests involve administration of the test chemical on multiple occasions. The test chemical may be administered one or more times each day, irregularly as when it is incorporated in the diet, at specific times such as during pregnancy, or in some cases regularly but only at weekly intervals. Also, in the prolonged test the experiment is usually conducted for not less than 90 days in the rat or mouse or a year in the dog. In contrast to the acute and prolonged types of test, the chronic toxicity tests are those in which the test chemical is administered for a substantial portion of the lifetime of the test animal. In the case of the mouse or rat, this is a period of 2 to 3 years. In the case of the dog, it is for 5 to 7 years.

Significant costs are incurred in establishing and maintaining large cohorts of test animals for such assays, especially the larger animals in chronic toxicity assays. Moreover, because of species specific effects, passing such toxicity tests does not ensure that a compound is free of toxic effects when used in humans. Such tests do, however, provide a standardized set of information for judging the safety of new compounds, and they provide a database for giving preliminary assessments of related compounds. An important area for improving toxicity determination would be the identification of new observables which are predictive of the outcome of the expensive and tedious animal assays.

In other medical fields, there has been significant interest in applying recent advances in biotechnology, particularly in DNA sequencing, to the identification and study of differentially expressed genes in healthy and diseased organisms, e.g. Adams et al, Science, 252: 1651-1656 (1991); Matsubara et al, Gene, 135: 265-274 (1993); Rosenberg et al, International patent application, PCT/US95/01863. The objectives of such applications include increasing our knowledge of disease processes, identifying genes that play important roles in the disease process, and providing diagnostic and therapeutic approaches that exploit the expressed genes or their

products. While such approaches are attractive, those based on exhaustive, or even sampled, sequencing of expressed genes are still beset by the enormous effort required: It is estimated that 30-35 thousand different genes are expressed in a typical mammalian tissue in any given state, e.g. Ausubel et al, Editors, Current Protocols, 5 5.8.1-5.8.4 (John Wiley & Sons, New York, 1992). Determining the sequences of even a small sample of that number of gene products is a major enterprise, requiring industrial-scale resources. Thus, the routine application of massive sequencing of expressed genes is still beyond current commercial technology.

The availability of new assays for assessing the toxicity of compounds, such 10 as candidate drugs, that would provide more comprehensive and precise information about the state of health of a test animal would be highly desirable. Such additional assays would preferably be less expensive, more rapid, and more convenient than current testing procedures, and would at the same time provide enough information to make early judgments regarding the safety of new compounds.

15

Summary of the Invention

An object of the invention is to provide a new approach to toxicity assessment based on an examination of gene expression patterns, or profiles, in in vitro or in vivo 20 test systems.

Another object of the invention is to provide a database on which to base 25 decisions concerning the toxicological properties of chemicals, particularly drug candidates.

A further object of the invention is to provide a method for analyzing gene expression patterns in selected tissues of test animals.

25 A still further object of the invention is to provide a system for identifying genes which are differentially expressed in response to exposure to a test compound.

Another object of the invention is to provide a rapid and reliable method for correlating gene expression with short term and long term toxicity in test animals.

30 Another object of the invention is to identify genes whose expression is predictive of deleterious toxicity.

The invention achieves these and other objects by providing a method for massively parallel signature sequencing of genes expressed in one or more selected tissues of an organism exposed to a test compound. An important feature of the invention is the application of novel DNA sorting and sequencing methodologies that 35 permit the formation of gene expression profiles for selected tissues by determining the sequence of portions of many thousands of different polynucleotides in parallel. Such profiles may be compared with those from tissues of control organisms at single or multiple time points to identify expression patterns predictive of toxicity.

The sorting methodology of the invention makes use of oligonucleotide tags that are members of a minimally cross-hybridizing set of oligonucleotides. The sequences of oligonucleotides of such a set differ from the sequences of every other member of the same set by at least two nucleotides. Thus, each member of such a set cannot form a duplex (or triplex) with the complement of any other member with less than two mismatches. Complements of oligonucleotide tags of the invention, referred to herein as "tag complements," may comprise natural nucleotides or non-natural nucleotide analogs. Preferably, tag complements are attached to solid phase supports. Such oligonucleotide tags when used with their corresponding tag complements provide a means of enhancing specificity of hybridization for sorting polynucleotides, such as cDNAs.

The polynucleotides to be sorted each have an oligonucleotide tag attached, such that different polynucleotides have different tags. As explained more fully below, this condition is achieved by employing a repertoire of tags substantially greater than the population of polynucleotides and by taking a sufficiently small sample of tagged polynucleotides from the full ensemble of tagged polynucleotides. After such sampling, when the populations of supports and polynucleotides are mixed under conditions which permit specific hybridization of the oligonucleotide tags with their respective complements, identical polynucleotides sort onto particular beads or regions. The sorted populations of polynucleotides can then be sequenced on the solid phase support by a "single-base" or "base-by-base" sequencing methodology, as described more fully below.

In one aspect, the method of the invention comprises the following steps: (a) administering the compound to a test organism; (b) extracting a population of mRNA molecules from each of one or more tissues of the test organism; (c) forming a separate population of cDNA molecules from each population of mRNA molecules extracted from the one or more tissues such that each cDNA molecule of the separate populations has an oligonucleotide tag attached, the oligonucleotide tags being selected from the same minimally cross-hybridizing set; (d) separately sampling each population of cDNA molecules such that substantially all different cDNA molecules within a separate population have different oligonucleotide tags attached; (e) sorting the cDNA molecules of each separate population by specifically hybridizing the oligonucleotide tags with their respective complements, the respective complements being attached as uniform populations of substantially identical complements in spatially discrete regions on one or more solid phase supports; (f) determining the nucleotide sequence of a portion of each of the sorted cDNA molecules of each separate population to form a frequency distribution of expressed genes for each of

the one or more tissues; and (g) correlating the frequency distribution of expressed genes in each of the one or more tissues with the toxicity of the compound.

An important aspect of the invention is the identification of genes whose expression is predictive of the toxicity of a compound. Once such genes are
5 identified, they may be employed in conventional assays, such as reverse transcriptase polymerase chain reaction (RT-PCR) assays for gene expression.

Brief Description of the Drawings

Figure 1 is a flow chart representation of an algorithm for generating
10 minimally cross-hybridizing sets of oligonucleotides.

Figure 2 diagrammatically illustrates an apparatus for carrying out polynucleotide sequencing in accordance with the invention.

Definitions

15 "Complement" or "tag complement" as used herein in reference to oligonucleotide tags refers to an oligonucleotide to which a oligonucleotide tag specifically hybridizes to form a perfectly matched duplex or triplex. In embodiments where specific hybridization results in a triplex, the oligonucleotide tag may be selected to be either double stranded or single stranded. Thus, where triplexes are
20 formed, the term "complement" is meant to encompass either a double stranded complement of a single stranded oligonucleotide tag or a single stranded complement of a double stranded oligonucleotide tag.

The term "oligonucleotide" as used herein includes linear oligomers of natural or modified monomers or linkages, including deoxyribonucleosides, ribonucleosides,
25 anomeric forms thereof, peptide nucleic acids (PNAs), and the like, capable of specifically binding to a target polynucleotide by way of a regular pattern of monomer-to-monomer interactions, such as Watson-Crick type of base pairing, base stacking, Hoogsteen or reverse Hoogsteen types of base pairing, or the like. Usually monomers are linked by phosphodiester bonds or analogs thereof to form
30 oligonucleotides ranging in size from a few monomeric units, e.g. 3-4, to several tens of monomeric units. Whenever an oligonucleotide is represented by a sequence of letters, such as "ATGCCTG," it will be understood that the nucleotides are in 5'→3' order from left to right and that "A" denotes deoxyadenosine, "C" denotes deoxycytidine, "G" denotes deoxyguanosine, and "T" denotes thymidine, unless
35 otherwise noted. Analogs of phosphodiester linkages include phosphorothioate, phosphorodithioate, phosphoranilidate, phosphoramidate, and the like. Usually oligonucleotides of the invention comprise the four natural nucleotides; however, they may also comprise non-natural nucleotide analogs. It is clear to those skilled in the

art when oligonucleotides having natural or non-natural nucleotides may be employed, e.g. where processing by enzymes is called for, usually oligonucleotides consisting of natural nucleotides are required.

5 "Perfectly matched" in reference to a duplex means that the poly- or oligonucleotide strands making up the duplex form a double stranded structure with one other such that every nucleotide in each strand undergoes Watson-Crick basepairing with a nucleotide in the other strand. The term also comprehends the pairing of nucleoside analogs, such as deoxyinosine, nucleosides with 2-aminopurine bases, and the like, that may be employed. In reference to a triplex, the term means
10 that the triplex consists of a perfectly matched duplex and a third strand in which every nucleotide undergoes Hoogsteen or reverse Hoogsteen association with a basepair of the perfectly matched duplex. Conversely, a "mismatch" in a duplex between a tag and an oligonucleotide means that a pair or triplet of nucleotides in the duplex or triplex fails to undergo Watson-Crick and/or Hoogsteen and/or reverse
15 Hoogsteen bonding.

As used herein, "nucleoside" includes the natural nucleosides, including 2'-deoxy and 2'-hydroxyl forms, e.g. as described in Kornberg and Baker, DNA Replication, 2nd Ed. (Freeman, San Francisco, 1992). "Analog" in reference to
nucleosides includes synthetic nucleosides having modified base moieties and/or
20 modified sugar moieties, e.g. described by Scheit, Nucleotide Analogs (John Wiley, New York, 1980); Uhlman and Peyman, Chemical Reviews, 90: 543-584 (1990). or the like, with the only proviso that they are capable of specific hybridization. Such analogs include synthetic nucleosides designed to enhance binding properties, reduce complexity, increase specificity, and the like.

25 As used herein "sequence determination" or "determining a nucleotide sequence" in reference to polynucleotides includes determination of partial as well as full sequence information of the polynucleotide. That is, the term includes sequence comparisons, fingerprinting, and like levels of information about a target polynucleotide, as well as the express identification and ordering of nucleosides,
30 usually each nucleoside, in a target polynucleotide. The term also includes the determination of the identification, ordering, and locations of one, two, or three of the four types of nucleotides within a target polynucleotide. For example, in some embodiments sequence determination may be effected by identifying the ordering and locations of a single type of nucleotide, e.g. cytosines, within the target polynucleotide
35 "CATCGC ..." so that its sequence is represented as a binary code, e.g. "100101 ..." for "C-(not C)-(not C)-C-(not C)-C ..." and the like.

As used herein, the term "complexity" in reference to a population of polynucleotides means the number of different species of molecule present in the population.

As used herein, the terms "gene expression profile," and "gene expression pattern" which is used equivalently, means a frequency distribution of sequences of portions of cDNA molecules sampled from a population of tag-cDNA conjugates. Generally, the portions of sequence are sufficiently long to uniquely identify the cDNA from which the portion arose. Preferably, the total number of sequences determined is at least 1000; more preferably, the total number of sequences
10 determined in a gene expression profile is at least ten thousand.

As used herein, "test organism" means any in vitro or in vivo system which provides measureable responses to exposure to test compounds. Typically, test organisms may be mammalian cell cultures, particularly of specific tissues, such as hepatocytes, neurons, kidney cells, colony forming cells, or the like, or test organisms
15 may be whole animals, such as rats, mice, hamsters, guinea pigs, dogs, cats, rabbits, pigs, monkeys, and the like.

Detailed Description of the Invention

The invention provides a method for determining the toxicity of a compound
20 by analyzing changes in the gene expression profiles in selected tissues of test organisms exposed to the compound. The invention also provides a method of identifying toxicity markers consisting of individual genes or a group of genes that is expressed acutely and which is correlated with prolonged or chronic toxicity, or suggests that the compound will have an undesirable cross reactivity. Gene
25 expression profiles are generated by sequencing portions of cDNA molecules construction from mRNA extracted from tissues of test organisms exposed to the compound being tested. As used herein, the term "tissue" is employed with its usual medical or biological meaning, except that in reference to an in vitro test system, such as a cell culture, it simply means a sample from the culture. Gene expression profiles
30 derived from test organisms are compared to gene expression profiles derived from control organisms to determine the genes which are differentially expressed in the test organism because of exposure to the compound being tested. In both cases, the sequence information of the gene expression profiles is obtained by massively parallel signature sequencing of cDNAs, which is implemented in steps (c) through (f) of the
35 above method.

Toxicity Assessment

Procedures for designing and conducting toxicity tests in in vitro and in vivo systems is well known, and is described in many texts on the subject, such as Loomis

et al. Loomis's Essentials of Toxicology, 4th Ed. (Academic Press, New York, 1996); Echobichon, The Basics of Toxicity Testing (CRC Press, Boca Raton, 1992); Frazier, editor, In Vitro Toxicity Testing (Marcel Dekker, New York, 1992); and the like.

5 In toxicity testing, two groups of test organisms are usually employed: one group serves as a control and the other group receives the test compound in a single dose (for acute toxicity tests) or a regimen of doses (for prolonged or chronic toxicity tests). Since in most cases, the extraction of tissue as called for in the method of the invention requires sacrificing the test animal, both the control group and the group receiving compound must be large enough to permit removal of animals for sampling
10 tissues, if it is desired to observe the dynamics of gene expression through the duration of an experiment.

In setting up a toxicity study, extensive guidance is provided in the literature for selecting the appropriate test organism for the compound being tested, route of administration, dose ranges, and the like. Water or physiological saline (0.9% NaCl
15 in water) is the solute of choice for the test compound since these solvents permit administration by a variety of routes. When this is not possible because of solubility limitations, it is necessary to resort to the use of vegetable oils such as corn oil or even organic solvents, of which propylene glycol is commonly used. Whenever possible the use of suspension or emulsion should be avoided except for oral
20 administration. Regardless of the route of administration, the volume required to administer a given dose is limited by the size of the animal that is used. It is desirable to keep the volume of each dose uniform within and between groups of animals. When rats or mice are used the volume administered by the oral route should not exceed 0.005 ml per gram of animal. Even when aqueous or physiological saline
25 solutions are used for parenteral injection the volumes that are tolerated are limited, although such solutions are ordinarily thought of as being innocuous. The intravenous LD₅₀ of distilled water in the mouse is approximately 0.044 ml per gram and that of isotonic saline is 0.068 ml per gram of mouse.

When a compound is to be administered by inhalation, special techniques for
30 generating test atmospheres are necessary. Dose estimation becomes very complicated. The methods usually involve aerosolization or nebulization of fluids containing the compound. If the agent to be tested is a fluid that has an appreciable vapor pressure, it may be administered by passing air through the solution under controlled temperature conditions. Under these conditions, dose is estimated from the
35 volume of air inhaled per unit time, the temperature of the solution, and the vapor pressure of the agent involved. Gases are metered from reservoirs. When particles of a solution are to be administered, unless the particle size is less than about 2 μ m the particles will not reach the terminal alveolar sacs in the lungs. A variety of

apparatuses and chambers are available to perform studies for detecting effects of irritant or other toxic endpoints when they are administered by inhalation. The preferred method of administering an agent to animals is via the oral route, either by intubation or by incorporating the agent in the feed.

5 Preferably, in designing a toxicity assessment, two or more species should be employed that handle the test compound as similarly to man as possible in terms of metabolism, absorption, excretion, tissue storage, and the like. Preferably, multiple doses or regimens at different concentrations should be employed to establish a dose-response relationship with respect to toxic effects. And preferably, the route of
10 administration to the test animal should be the same as, or as similar as possible to, the route of administration of the compound to man. Effects obtained by one route of administration to test animals are not a priori applicable to effects by another route of administration to man. For example, food additives for man should be tested by admixture of the material in the diet of the test animals.

15 Acute toxicity tests consist of administering a compound to test organisms on one occasion. The purpose of such test is to determine the symptomatology consequent to administration of the compound and to determine the degree of lethality of the compound. The initial procedure is to perform a series of range-finding doses of the compound in a single species. This necessitates selection of a route of
20 administration, preparation of the compound in a form suitable for administration by the selected route, and selection of an appropriate species. Preferably, initial acute toxicity studies are performed on either rats or mice because of their low cost, their availability, and the availability of abundant toxicologic reference data on these species. Prolonged toxicity tests consist of administering a compound to test
25 organisms repeatedly, usually on a daily basis, over a period of 3 to 4 months. Two practical factors are encountered that place constraints on the design of such tests: First, the available routes of administration are limited because the route selected must be suitable for repeated administration without inducing harmful effects. And second, blood, urine, and perhaps other samples, should be taken repeatedly without
30 inducing significant harm to the test animals. Preferably, in the method of the invention the gene expression profiles are obtained in conjunction with the measurement of the traditional toxicologic parameters, such as listed in the table below:

35

Hematology	Blood Chemistry	Urine Analyses
erythrocyte count	sodium	pH
total leukocyte count	potassium	specific gravity
differential leukocyte count	chloride	total protein
hematocrit	calcium	sediment
hemoglobin	carbon dioxide	glucose
	serum glutamine-pyruvate transaminase	ketones
	serum glutamin-oxalacetic transaminase	bilirubin
	serum protein electrophoresis	
	blood sugar	
	blood urea nitrogen	
	total serum protein	
	serum albumin	
	total serum bilirubin	

5

Oligonucleotide Tags and Tag Complements

Oligonucleotide tags are members of a minimally cross-hybridizing set of oligonucleotides. The sequences of oligonucleotides of such a set differ from the sequences of every other member of the same set by at least two nucleotides. Thus, each member of such a set cannot form a duplex (or triplex) with the complement of any other member with less than two mismatches. Complements of oligonucleotide tags, referred to herein as "tag complements," may comprise natural nucleotides or non-natural nucleotide analogs. Preferably, tag complements are attached to solid phase supports. Such oligonucleotide tags when used with their corresponding tag complements provide a means of enhancing specificity of hybridization for sorting, tracking, or labeling molecules, especially polynucleotides.

Minimally cross-hybridizing sets of oligonucleotide tags and tag complements may be synthesized either combinatorially or individually depending on the size of the set desired and the degree to which cross-hybridization is sought to be minimized (or stated another way, the degree to which specificity is sought to be enhanced). For example, a minimally cross-hybridizing set may consist of a set of individually synthesized 10-mer sequences that differ from each other by at least 4 nucleotides, such set having a maximum size of 332 (when composed of 3 kinds of nucleotides and counted using a computer program such as disclosed in Appendix Ic). Alternatively, a minimally cross-hybridizing set of oligonucleotide tags may also be

assembled combinatorially from subunits which themselves are selected from a minimally cross-hybridizing set. For example, a set of minimally cross-hybridizing 12-mers differing from one another by at least three nucleotides may be synthesized by assembling 3 subunits selected from a set of minimally cross-hybridizing 4-mers that each differ from one another by three nucleotides. Such an embodiment gives a maximally sized set of 9^3 , or 729, 12-mers. The number 9 is number of oligonucleotides listed by the computer program of Appendix Ia, which assumes, as with the 10-mers, that only 3 of the 4 different types of nucleotides are used. The set is described as "maximal" because the computer programs of Appendices Ia-c provide the largest set for a given input (e.g. length, composition, difference in number of nucleotides between members). Additional minimally cross-hybridizing sets may be formed from subsets of such calculated sets.

Oligonucleotide tags may be single stranded and be designed for specific hybridization to single stranded tag complements by duplex formation or for specific hybridization to double stranded tag complements by triplex formation. Oligonucleotide tags may also be double stranded and be designed for specific hybridization to single stranded tag complements by triplex formation.

When synthesized combinatorially, an oligonucleotide tag preferably consists of a plurality of subunits, each subunit consisting of an oligonucleotide of 3 to 9 nucleotides in length wherein each subunit is selected from the same minimally cross-hybridizing set. In such embodiments, the number of oligonucleotide tags available depends on the number of subunits per tag and on the length of the subunits. The number is generally much less than the number of all possible sequences the length of the tag, which for a tag n nucleotides long would be 4^n .

Complements of oligonucleotide tags attached to a solid phase support are used to sort polynucleotides from a mixture of polynucleotides each containing a tag. Complements of the oligonucleotide tags are synthesized on the surface of a solid phase support, such as a microscopic bead or a specific location on an array of synthesis locations on a single support, such that populations of identical sequences are produced in specific regions. That is, the surface of each support, in the case of a bead, or of each region, in the case of an array, is derivatized by only one type of complement which has a particular sequence. The population of such beads or regions contains a repertoire of complements with distinct sequences. As used herein in reference to oligonucleotide tags and tag complements, the term "repertoire" means the set of minimally cross-hybridizing set of oligonucleotides that make up the tags in a particular embodiment or the corresponding set of tag complements.

The polynucleotides to be sorted each have an oligonucleotide tag attached, such that different polynucleotides have different tags. As explained more fully

below, this condition is achieved by employing a repertoire of tags substantially greater than the population of polynucleotides and by taking a sufficiently small sample of tagged polynucleotides from the full ensemble of tagged polynucleotides. After such sampling, when the populations of supports and polynucleotides are mixed
 5 under conditions which permit specific hybridization of the oligonucleotide tags with their respective complements, identical polynucleotides sort onto particular beads or regions.

The nucleotide sequences of oligonucleotides of a minimally cross-hybridizing set are conveniently enumerated by simple computer programs, such as those
 10 exemplified by programs whose source codes are listed in Appendices Ia and Ib. Program minhx of Appendix Ia computes all minimally cross-hybridizing sets having 4-mer subunits composed of three kinds of nucleotides. Program tagN of Appendix Ib enumerates longer oligonucleotides of a minimally cross-hybridizing set. Similar algorithms and computer programs are readily written for listing oligonucleotides of
 15 minimally cross-hybridizing sets for any embodiment of the invention. Table I below provides guidance as to the size of sets of minimally cross-hybridizing oligonucleotides for the indicated lengths and number of nucleotide differences. The above computer programs were used to generate the numbers.

20

Table I

Oligonucleotide Word Length	Nucleotide Difference between Oligonucleotides of Minimally Cross- Hybridizing Set	Maximal Size of Minimally Cross- Hybridizing Set	Size of Repertoire with Four Words	Size of Repertoire with Five Words
4	3	9	6561	5.90×10^4
6	3	27	5.3×10^5	1.43×10^7
7	4	27	5.3×10^5	1.43×10^7
7	5	8	4096	3.28×10^4
8	3	190	1.30×10^9	2.48×10^{11}
8	4	62	1.48×10^7	9.16×10^8
8	5	18	1.05×10^5	1.89×10^6
9	5	39	2.31×10^6	9.02×10^7
10	5	332	1.21×10^{10}	
10	6	28	6.15×10^5	1.72×10^7
11	5	187		
18	6	≈ 25000		

18

12

24

For some embodiments of the invention, where extremely large repertoires of tags are not required, oligonucleotide tags of a minimally cross-hybridizing set may be separately synthesized. Sets containing several hundred to several thousands, or
5 even several tens of thousands, of oligonucleotides may be synthesized directly by a variety of parallel synthesis approaches, e.g. as disclosed in Frank et al, U.S. patent 4,689,405; Frank et al, Nucleic Acids Research, 11: 4365-4377 (1983); Matson et al, Anal. Biochem., 224: 110-116 (1995); Fodor et al, International application PCT/US93/04145; Pease et al, Proc. Natl. Acad. Sci., 91: 5022-5026 (1994);
10 Southern et al, J. Biotechnology, 35: 217-227 (1994), Brennan, International application PCT/US94/05896; Lashkari et al, Proc. Natl. Acad. Sci., 92: 7912-7915 (1995); or the like.

Preferably, oligonucleotide tags of the invention are synthesized combinatorially out of subunits between three and six nucleotides in length and
15 selected from the same minimally cross-hybridizing set. For oligonucleotides in this range, the members of such sets may be enumerated by computer programs based on the algorithm of Fig. 1.

The algorithm of Fig. 1 is implemented by first defining the characteristics of the subunits of the minimally cross-hybridizing set, i.e. length, number of base
20 differences between members, and composition, e.g. do they consist of two, three, or four kinds of bases. A table M_n , $n=1$, is generated (100) that consists of all possible sequences of a given length and composition. An initial subunit S_1 is selected and compared (120) with successive subunits S_i for $i=n+1$ to the end of the table. Whenever a successive subunit has the required number of mismatches to be a
25 member of the minimally cross-hybridizing set, it is saved in a new table M_{n+1} (125), that also contains subunits previously selected in prior passes through step 120. For example, in the first set of comparisons, M_2 will contain S_1 ; in the second set of comparisons, M_3 will contain S_1 and S_2 ; in the third set of comparisons, M_4 will contain S_1 , S_2 , and S_3 ; and so on. Similarly, comparisons in table M_j will be
30 between S_j and all successive subunits in M_j . Note that each successive table M_{n+1} is smaller than its predecessors as subunits are eliminated in successive passes through step 130. After every subunit of table M_n has been compared (140) the old table is replaced by the new table M_{n+1} , and the next round of comparisons are begun. The process stops (160) when a table M_n is reached that contains no
35 successive subunits to compare to the selected subunit S_i , i.e. $M_n=M_{n+1}$.

Preferably, minimally cross-hybridizing sets comprise subunits that make approximately equivalent contributions to duplex stability as every other subunit in

the set. In this way, the stability of perfectly matched duplexes between every subunit and its complement is approximately equal. Guidance for selecting such sets is provided by published techniques for selecting optimal PCR primers and calculating duplex stabilities, e.g. Rychlik et al, Nucleic Acids Research, 17: 8543-8551 (1989) and 18: 6409-6412 (1990); Breslauer et al, Proc. Natl. Acad. Sci., 83: 3746-3750 (1986); Wetmur, Crit. Rev. Biochem. Mol. Biol., 26: 227-259 (1991); and the like. For shorter tags, e.g. about 30 nucleotides or less, the algorithm described by Rychlik and Wetmur is preferred, and for longer tags, e.g. about 30-35 nucleotides or greater, an algorithm disclosed by Suggs et al, pages 683-693 in Brown, editor, ICN-UCLA Symp. Dev. Biol., Vol. 23 (Academic Press, New York, 1981) may be conveniently employed. Clearly, there are many approaches available to one skilled in the art for designing sets of minimally cross-hybridizing subunits within the scope of the invention. For example, to minimize the effects of different base-stacking energies of terminal nucleotides when subunits are assembled, subunits may be provided that have the same terminal nucleotides. In this way, when subunits are linked, the sum of the base-stacking energies of all the adjoining terminal nucleotides will be the same, thereby reducing or eliminating variability in tag melting temperatures.

A "word" of terminal nucleotides, shown in *italic* below, may also be added to each end of a tag so that a perfect match is always formed between it and a similar terminal "word" on any other tag complement. Such an augmented tag would have the form:

<i>W</i>	<i>W</i> ₁	<i>W</i> ₂	...	<i>W</i> _{k-1}	<i>W</i> _k	<i>W</i>
<i>W'</i>	<i>W</i> ₁ '	<i>W</i> ₂ '	...	<i>W</i> _{k-1} '	<i>W</i> _k '	<i>W'</i>

where the primed *W*'s indicate complements. With ends of tags always forming perfectly matched duplexes, all mismatched words will be internal mismatches thereby reducing the stability of tag-complement duplexes that otherwise would have mismatched words at their ends. It is well known that duplexes with internal mismatches are significantly less stable than duplexes with the same mismatch at a terminus.

A preferred embodiment of minimally cross-hybridizing sets are those whose subunits are made up of three of the four natural nucleotides. As will be discussed more fully below, the absence of one type of nucleotide in the oligonucleotide tags permits target polynucleotides to be loaded onto solid phase supports by use of the 5'→3' exonuclease activity of a DNA polymerase. The following is an exemplary minimally cross-hybridizing set of subunits each comprising four nucleotides selected from the group consisting of A, G, and T:

5

Table II

Word:	w ₁	w ₂	w ₃	w ₄
Sequence:	GATT	TGAT	TAGA	TTTG
Word:	w ₅	w ₆	w ₇	w ₈
Sequence:	GTAA	AGTA	ATGT	AAAG

10 In this set, each member would form a duplex having three mismatched bases with the complement of every other member.

Further exemplary minimally cross-hybridizing sets are listed below in Table III. Clearly, additional sets can be generated by substituting different groups of nucleotides, or by using subsets of known minimally cross-hybridizing sets.

15

Table III

Exemplary Minimally Cross-Hybridizing Sets of 4-mer Subunits

<u>Set 1</u>	<u>Set 2</u>	<u>Set 3</u>	<u>Set 4</u>	<u>Set 5</u>	<u>Set 6</u>
CATT	ACCC	AAAC	AAAG	AACA	AACG
CTAA	AGGG	ACCA	ACCA	ACAC	ACAA
TCAT	CACG	AGGG	AGGC	AGGG	AGGC
ACTA	CCGA	CACG	CACC	CAAG	CAAC
TACA	CGAC	CCGC	CCGG	CCGC	CCGG
TTTC	GAGC	CGAA	CGAA	CGCA	CGCA
ATCT	GCAG	GAGA	GAGA	GAGA	GAGA
AAAC	GGCA	GCAG	GCAC	GCCG	GCCC
	AAAA	GGCC	GGCG	GGAC	GGAG

<u>Set 7</u>	<u>Set 8</u>	<u>Set 9</u>	<u>Set 10</u>	<u>Set 11</u>	<u>Set 12</u>
AAGA	AAGC	AAGG	ACAG	ACCG	ACGA
ACAC	ACAA	ACAA	AACA	AAAA	AAAC
AGCG	AGCG	AGCC	AGGC	AGGC	AGCG
CAAG	CAAG	CAAC	CAAC	CACC	CACA
CCCA	CCCC	CCCG	CCGA	CCGA	CCAG
CGGC	CGGA	CGGA	CGCG	CGAG	CGGC
GACC	GACA	GACA	GAGG	GAGG	GAGG
GCGG	GCGG	GCGC	GCCC	GCAC	GCCC
GGAA	GGAC	GGAG	GGAA	GGCA	GGAA

The oligonucleotide tags of the invention and their complements are conveniently synthesized on an automated DNA synthesizer, e.g. an Applied Biosystems, Inc. (Foster City, California) model 392 or 394 DNA/RNA Synthesizer, using standard chemistries, such as phosphoramidite chemistry, e.g. disclosed in the following references: Beaucage and Iyer, Tetrahedron, 48: 2223-2311 (1992); Molko et al, U.S. patent 4,980,460; Koster et al, U.S. patent 4,725,677; Caruthers et al, U.S. patents 4,415,732; 4,458,066; and 4,973,679; and the like. Alternative chemistries, e.g. resulting in non-natural backbone groups, such as phosphorothioate, phosphoramidate, and the like, may also be employed provided that the resulting oligonucleotides are capable of specific hybridization. In some embodiments, tags may comprise naturally occurring nucleotides that permit processing or manipulation by enzymes, while the corresponding tag complements may comprise non-natural nucleotide analogs, such as peptide nucleic acids, or like compounds, that promote the formation of more stable duplexes during sorting.

When microparticles are used as supports, repertoires of oligonucleotide tags and tag complements may be generated by subunit-wise synthesis via "split and mix" techniques, e.g. as disclosed in Shortle et al, International patent application PCT/US93/03418 or Lytle et al, Biotechniques, 19: 274-280 (1995). Briefly, the basic unit of the synthesis is a subunit of the oligonucleotide tag. Preferably, phosphoramidite chemistry is used and 3' phosphoramidite oligonucleotides are prepared for each subunit in a minimally cross-hybridizing set, e.g. for the set first listed above, there would be eight 4-mer 3'-phosphoramidites. Synthesis proceeds as disclosed by Shortle et al or in direct analogy with the techniques employed to generate diverse oligonucleotide libraries using nucleosidic monomers, e.g. as disclosed in Telenius et al, Genomics, 13: 718-725 (1992); Welsh et al, Nucleic Acids Research, 19: 5275-5279 (1991); Grothues et al, Nucleic Acids Research, 21: 1321-1322 (1993); Hartley, European patent application 90304496.4; Lam et al, Nature, 354: 82-84 (1991); Zuckerman et al, Int. J. Pept. Protein Research, 40: 498-507 (1992); and the like. Generally, these techniques simply call for the application of

mixtures of the activated monomers to the growing oligonucleotide during the coupling steps. Preferably, oligonucleotide tags and tag complements are synthesized on a DNA synthesizer having a number of synthesis chambers which is greater than or equal to the number of different kinds of words used in the construction of the tags.

- 5 That is, preferably there is a synthesis chamber corresponding to each type of word. In this embodiment, words are added nucleotide-by-nucleotide, such that if a word consists of five nucleotides there are five monomer couplings in each synthesis chamber. After a word is completely synthesized, the synthesis supports are removed from the chambers, mixed, and redistributed back to the chambers for the next cycle
10 of word addition. This latter embodiment takes advantage of the high coupling yields of monomer addition, e.g. in phosphoramidite chemistries.

- Double stranded forms of tags may be made by separately synthesizing the complementary strands followed by mixing under conditions that permit duplex formation. Alternatively, double stranded tags may be formed by first synthesizing a
15 single stranded repertoire linked to a known oligonucleotide sequence that serves as a primer binding site. The second strand is then synthesized by combining the single stranded repertoire with a primer and extending with a polymerase. This latter approach is described in Oliphant et al, Gene, 44: 177-183 (1986). Such duplex tags may then be inserted into cloning vectors along with target polynucleotides for sorting
20 and manipulation of the target polynucleotide in accordance with the invention.

- When tag complements are employed that are made up of nucleotides that have enhanced binding characteristics, such as PNAs or oligonucleotide N3'→P5' phosphoramidates, sorting can be implemented through the formation of D-loops between tags comprising natural nucleotides and their PNA or phosphoramidate
25 complements, as an alternative to the "stripping" reaction employing the 3'→5' exonuclease activity of a DNA polymerase to render a tag single stranded.

- Oligonucleotide tags of the invention may range in length from 12 to 60 nucleotides or basepairs. Preferably, oligonucleotide tags range in length from 18 to 40 nucleotides or basepairs. More preferably, oligonucleotide tags range in length
30 from 25 to 40 nucleotides or basepairs. In terms of preferred and more preferred numbers of subunits, these ranges may be expressed as follows:

Table IV
Numbers of Subunits in Tags in Preferred Embodiments

35

Monomers
in Subunit

<u>Nucleotides in Oligonucleotide Tag</u>		
(12-60)	(18-40)	(25-40)

3	4-20 subunits	6-13 subunits	8-13 subunits
4	3-15 subunits	4-10 subunits	6-10 subunits
5	2-12 subunits	3-8 subunits	5-8 subunits
6	2-10 subunits	3-6 subunits	4-6 subunits

Most preferably, oligonucleotide tags are single stranded and specific hybridization occurs via Watson-Crick pairing with a tag complement.

Preferably, repertoires of single stranded oligonucleotide tags of the invention contain at least 100 members; more preferably, repertoires of such tags contain at least 1000 members; and most preferably, repertoires of such tags contain at least 10,000 members.

Triplex Tags

In embodiments where specific hybridization occurs via triplex formation, coding of tag sequences follows the same principles as for duplex-forming tags; however, there are further constraints on the selection of subunit sequences. Generally, third strand association via Hoogsteen type of binding is most stable along homopyrimidine-homopurine tracks in a double stranded target. Usually, base triplets form in T-A*T or C-G*C motifs (where "-" indicates Watson-Crick pairing and "*" indicates Hoogsteen type of binding); however, other motifs are also possible. For example, Hoogsteen base pairing permits parallel and antiparallel orientations between the third strand (the Hoogsteen strand) and the purine-rich strand of the duplex to which the third strand binds, depending on conditions and the composition of the strands. There is extensive guidance in the literature for selecting appropriate sequences, orientation, conditions, nucleoside type (e.g. whether ribose or deoxyribose nucleosides are employed), base modifications (e.g. methylated cytosine, and the like) in order to maximize, or otherwise regulate, triplex stability as desired in particular embodiments, e.g. Roberts et al, Proc. Natl. Acad. Sci., 88: 9397-9401 (1991); Roberts et al, Science, 258: 1463-1466 (1992); Roberts et al, Proc. Natl. Acad. Sci., 93: 4320-4325 (1996); Distefano et al, Proc. Natl. Acad. Sci., 90: 1179-1183 (1993); Mergny et al, Biochemistry, 30: 9791-9798 (1991); Cheng et al, J. Am. Chem. Soc., 114: 4465-4474 (1992); Beal and Dervan, Nucleic Acids Research, 20: 2773-2776 (1992); Beal and Dervan, J. Am. Chem. Soc., 114: 4976-4982 (1992); Giovannangeli et al, Proc. Natl. Acad. Sci., 89: 8631-8635 (1992); Moser and Dervan, Science, 238: 645-650 (1987); McShan et al, J. Biol. Chem., 267: 5712-5721 (1992); Yoon et al, Proc. Natl. Acad. Sci., 89: 3840-3844 (1992); Blume et al, Nucleic Acids Research, 20: 1777-1784 (1992); Thuong and Helene, Angew. Chem. Int. Ed. Engl.

32: 666-690 (1993); Escude et al, Proc. Natl. Acad. Sci., 93: 4365-4369 (1996); and the like. Conditions for annealing single-stranded or duplex tags to their single-stranded or duplex complements are well known, e.g. Ji et al, Anal. Chem. 65: 1323-1328 (1993); Cantor et al, U.S. patent 5,482,836; and the like. Use of triplex tags has the advantage of not requiring a "stripping" reaction with polymerase to expose the tag for annealing to its complement.

Preferably, oligonucleotide tags of the invention employing triplex hybridization are double stranded DNA and the corresponding tag complements are single stranded. More preferably, 5-methylcytosine is used in place of cytosine in the tag complements in order to broaden the range of pH stability of the triplex formed between a tag and its complement. Preferred conditions for forming triplexes are fully disclosed in the above references. Briefly, hybridization takes place in concentrated salt solution, e.g. 1.0 M NaCl, 1.0 M potassium acetate, or the like, at pH below 5.5 (or 6.5 if 5-methylcytosine is employed). Hybridization temperature depends on the length and composition of the tag; however, for an 18-20-mer tag of longer, hybridization at room temperature is adequate. Washes may be conducted with less concentrated salt solutions, e.g. 10 mM sodium acetate, 100 mM MgCl₂, pH 5.8, at room temperature. Tags may be eluted from their tag complements by incubation in a similar salt solution at pH 9.0.

Minimally cross-hybridizing sets of oligonucleotide tags that form triplexes may be generated by the computer program of Appendix Ic, or similar programs. An exemplary set of double stranded 8-mer words are listed below in capital letters with the corresponding complements in small letters. Each such word differs from each of the other words in the set by three base pairs.

Table V
Exemplary Minimally Cross-Hybridizing
Set of DoubleStranded 8-mer Tags

5' -AAGGAGAG	5' -AAAGGGGA	5' -AGAGAAGA	5' -AGGGGGGG
3' -TTCCTCTC	3' -TTTCCCCT	3' -TCTCTTCT	3' -TCCCCCCC
3' -ttcctctc	3' -tttcccct	3' -tctcttct	3' -tccccccc
5' -AAAAAATA	5' -AAGAGAGA	5' -AGGAAAAG	5' -GAAAGGAG
3' -TTTTTTTT	3' -TTCTCTCT	3' -TCCTTTTC	3' -CTTTCCTC
3' -tttttttt	3' -ttctctct	3' -tccttttc	3' -ctttcctc
5' -AAAAAGGG	5' -AGAAGAGG	5' -AGGAAGGA	5' -GAAGAAGG
3' -TTTTTCCC	3' -TCTTCTCC	3' -TCCTTCCT	3' -CTTCTTCC
3' -tttttccc	3' -tcttctcc	3' -tccttcct	3' -cttcttcc
5' -AAAGGAAG	5' -AGAAGGAA	5' -AGGGGAAA	5' -GAAGAGAA
3' -TTTCCTTC	3' -TCTTCCTT	3' -TCCCCTTT	3' -CTTCTCTT
3' -tttccttc	3' -tcttcctt	3' -tccccttt	3' -cttctctt

5

10

Table VI
Repertoire Size of Various Double Stranded Tags
That Form Triplexes with Their Tag Complements

Oligonucleotide Word Length	Nucleotide Difference between Oligonucleotides of Minimally Cross- Hybridizing Set	Maximal Size of Minimally Cross- Hybridizing Set	Size of Repertoire with Four Words	Size of Repertoire with Five Words
4	2	8	4096	3.2×10^4
6	3	8	4096	3.2×10^4
8	3	16	6.5×10^4	1.05×10^6
10	5	8	4096	
15	5	92		
20	6	765		
20	8	92		
20	10	22		

15 Preferably, repertoires of double stranded oligonucleotide tags of the invention contain at least 10 members; more preferably, repertoires of such tags contain at least 100 members. Preferably, words are between 4 and 8 nucleotides in length for combinatorially synthesized double stranded oligonucleotide tags, and oligonucleotide tags are between 12 and 60 base pairs in length. More preferably, such tags are
20 between 18 and 40 base pairs in length.

Solid Phase Supports

Solid phase supports for use with the invention may have a wide variety of forms, including microparticles, beads, and membranes, slides, plates, micromachined
25 chips, and the like. Likewise, solid phase supports of the invention may comprise a

wide variety of compositions, including glass, plastic, silicon, alkanethiolate-derivatized gold, cellulose, low cross-linked and high cross-linked polystyrene, silica gel, polyamide, and the like. Preferably, either a population of discrete particles are employed such that each has a uniform coating, or population, of complementary sequences of the same tag (and no other), or a single or a few supports are employed with spatially discrete regions each containing a uniform coating, or population, of complementary sequences to the same tag (and no other). In the latter embodiment, the area of the regions may vary according to particular applications; usually, the regions range in area from several μm^2 , e.g. 3-5, to several hundred μm^2 , e.g. 100-500. Preferably, such regions are spatially discrete so that signals generated by events, e.g. fluorescent emissions, at adjacent regions can be resolved by the detection system being employed. In some applications, it may be desirable to have regions with uniform coatings of more than one tag complement, e.g. for simultaneous sequence analysis, or for bringing separately tagged molecules into close proximity.

Tag complements may be used with the solid phase support that they are synthesized on, or they may be separately synthesized and attached to a solid phase support for use, e.g. as disclosed by Lund et al, *Nucleic Acids Research*, 16: 10861-10880 (1988); Albretsen et al, *Anal. Biochem.*, 189: 40-50 (1990); Wolf et al, *Nucleic Acids Research*, 15: 2911-2926 (1987); or Ghosh et al, *Nucleic Acids Research*, 15: 5353-5372 (1987). Preferably, tag complements are synthesized on and used with the same solid phase support, which may comprise a variety of forms and include a variety of linking moieties. Such supports may comprise microparticles or arrays, or matrices, of regions where uniform populations of tag complements are synthesized. A wide variety of microparticle supports may be used with the invention, including microparticles made of controlled pore glass (CPG), highly cross-linked polystyrene, acrylic copolymers, cellulose, nylon, dextran, latex, polyacrolein, and the like, disclosed in the following exemplary references: *Meth. Enzymol.*, Section A, pages 11-147, vol. 44 (Academic Press, New York, 1976); U.S. patents 4,678,814; 4,413,070; and 4,046,720; and Pon, Chapter 19, in Agrawal, editor, *Methods in Molecular Biology*, Vol. 20, (Humana Press, Totowa, NJ, 1993). Microparticle supports further include commercially available nucleoside-derivatized CPG and polystyrene beads (e.g. available from Applied Biosystems, Foster City, CA); derivatized magnetic beads; polystyrene grafted with polyethylene glycol (e.g., TentaGelTM, Rapp Polymere, Tubingen Germany); and the like. Selection of the support characteristics, such as material, porosity, size, shape, and the like, and the type of linking moiety employed depends on the conditions under which the tags are used. For example, in applications involving successive processing with enzymes, supports and linkers that minimize steric hindrance of the enzymes and that facilitate

access to substrate are preferred. Other important factors to be considered in selecting the most appropriate microparticle support include size uniformity, efficiency as a synthesis support, degree to which surface area known, and optical properties, e.g. as explain more fully below, clear smooth beads provide instrumental advantages when handling large numbers of beads on a surface.

Exemplary linking moieties for attaching and/or synthesizing tags on microparticle surfaces are disclosed in Pon et al, *Biotechniques*, 6:768-775 (1988); Webb, U.S. patent 4,659,774; Barany et al, International patent application PCT/US91/06103; Brown et al, *J. Chem. Soc. Commun.*, 1989: 891-893; Damha et al, *Nucleic Acids Research*, 18: 3813-3821 (1990); Beattie et al, *Clinical Chemistry*, 39: 719-722 (1993); Maskos and Southern, *Nucleic Acids Research*, 20: 1679-1684 (1992); and the like.

As mentioned above, tag complements may also be synthesized on a single (or a few) solid phase support to form an array of regions uniformly coated with tag complements. That is, within each region in such an array the same tag complement is synthesized. Techniques for synthesizing such arrays are disclosed in McGall et al, International application PCT/US93/03767; Pease et al, *Proc. Natl. Acad. Sci.*, 91: 5022-5026 (1994); Southern and Maskos, International application PCT/GB89/01114; Maskos and Southern (cited above); Southern et al, *Genomics*, 13: 1008-1017 (1992); and Maskos and Southern, *Nucleic Acids Research*, 21: 4663-4669 (1993).

Preferably, the invention is implemented with microparticles or beads uniformly coated with complements of the same tag sequence. Microparticle supports and methods of covalently or noncovalently linking oligonucleotides to their surfaces are well known, as exemplified by the following references: Beaucage and Iyer (cited above); Gait, editor, *Oligonucleotide Synthesis: A Practical Approach* (IRL Press, Oxford, 1984); and the references cited above. Generally, the size and shape of a microparticle is not critical; however, microparticles in the size range of a few, e.g. 1-2, to several hundred, e.g. 200-1000 μm diameter are preferable, as they facilitate the construction and manipulation of large repertoires of oligonucleotide tags with minimal reagent and sample usage.

In some preferred applications, commercially available controlled-pore glass (CPG) or polystyrene supports are employed as solid phase supports in the invention. Such supports come available with base-labile linkers and initial nucleosides attached. e.g. Applied Biosystems (Foster City, CA). Preferably, microparticles having pore size between 500 and 1000 angstroms are employed.

In other preferred applications, non-porous microparticles are employed for their optical properties, which may be advantageously used when tracking large

numbers of microparticles on planar supports, such as a microscope slide. Particularly preferred non-porous microparticles are the glycidal methacrylate (GMA) beads available from Bangs Laboratories (Carmel, IN). Such microparticles are useful in a variety of sizes and derivatized with a variety of linkage groups for synthesizing tags or tag complements. Preferably, for massively parallel manipulations of tagged microparticles, 5 μ m diameter GMA beads are employed.

10

Attaching Tags to Polynucleotides
For Sorting onto Solid Phase Supports

An important aspect of the invention is the sorting and attachment of a populations of polynucleotides, e.g. from a cDNA library, to microparticles or to separate regions on a solid phase support such that each microparticle or region has substantially only one kind of polynucleotide attached. This objective is accomplished by insuring that substantially all different polynucleotides have different tags attached. This condition, in turn, is brought about by taking a sample of the full ensemble of tag-polynucleotide conjugates for analysis. (It is acceptable that identical polynucleotides have different tags, as it merely results in the same polynucleotide being operated on or analyzed twice in two different locations.) Such sampling can be carried out either overtly--for example, by taking a small volume from a larger mixture--after the tags have been attached to the polynucleotides, it can be carried out inherently as a secondary effect of the techniques used to process the polynucleotides and tags, or sampling can be carried out both overtly and as an inherent part of processing steps.

Preferably, in constructing a cDNA library where substantially all different cDNAs have different tags, a tag repertoire is employed whose complexity, or number of distinct tags, greatly exceeds the total number of mRNAs extracted from a cell or tissue sample. Preferably, the complexity of the tag repertoire is at least 10 times that of the polynucleotide population; and more preferably, the complexity of the tag repertoire is at least 100 times that of the polynucleotide population. Below, a protocol is disclosed for cDNA library construction using a primer mixture that contains a full repertoire of exemplary 9-word tags. Such a mixture of tag-containing primers has a complexity of 8^9 , or about 1.34×10^8 . As indicated by Winslow et al, Nucleic Acids Research, 19: 3251-3253 (1991), mRNA for library construction can be extracted from as few as 10-100 mammalian cells. Since a single mammalian cell contains about 5×10^5 copies of mRNA molecules of about 3.4×10^4 different kinds,

by standard techniques one can isolate the mRNA from about 100 cells, or (theoretically) about 5×10^7 mRNA molecules. Comparing this number to the complexity of the primer mixture shows that without any additional steps, and even assuming that mRNAs are converted into cDNAs with perfect efficiency (1% efficiency or less is more accurate), the cDNA library construction protocol results in a population containing no more than 37% of the total number of different tags. That is, without any overt sampling step at all, the protocol inherently generates a sample that comprises 37%, or less, of the tag repertoire. The probability of obtaining a double under these conditions is about 5%, which is within the preferred range. With mRNA from 10 cells, the fraction of the tag repertoire sampled is reduced to only 3.7%, even assuming that all the processing steps take place at 100% efficiency. In fact, the efficiencies of the processing steps for constructing cDNA libraries are very low, a "rule of thumb" being that good library should contain about 10^8 cDNA clones from mRNA extracted from 10^6 mammalian cells.

Use of larger amounts of mRNA in the above protocol, or for larger amounts of polynucleotides in general, where the number of such molecules exceeds the complexity of the tag repertoire, a tag-polynucleotide conjugate mixture potentially contains every possible pairing of tags and types of mRNA or polynucleotide. In such cases, overt sampling may be implemented by removing a sample volume after a serial dilution of the starting mixture of tag-polynucleotide conjugates. The amount of dilution required depends on the amount of starting material and the efficiencies of the processing steps, which are readily estimated.

If mRNA were extracted from 10^6 cells (which would correspond to about 0.5 μg of poly(A)⁺ RNA), and if primers were present in about 10-100 fold concentration excess--as is called for in a typical protocol, e.g. Sambrook et al, Molecular Cloning, Second Edition, page 8.61 [10 μL 1.8 kb mRNA at 1 mg/mL equals about 1.68×10^{-11} moles and 10 μL 18-mer primer at 1 mg/mL equals about 1.68×10^{-9} moles], then the total number of tag-polynucleotide conjugates in a cDNA library would simply be equal to or less than the starting number of mRNAs, or about 5×10^{11} vectors containing tag-polynucleotide conjugates--again this assumes that each step in cDNA construction--first strand synthesis, second strand synthesis, ligation into a vector--occurs with perfect efficiency, which is a very conservative estimate. The actual number is significantly less.

If a sample of n tag-polynucleotide conjugates are randomly drawn from a reaction mixture--as could be effected by taking a sample volume, the probability of drawing conjugates having the same tag is described by the Poisson distribution, $P(r) = e^{-\lambda} (\lambda)^r / r!$, where r is the number of conjugates having the same tag and $\lambda = np$, where p is the probability of a given tag being selected. If $n = 10^6$ and $p = 1/(1.34 \times$

10⁸), then $\lambda = .00746$ and $P(2) = 2.76 \times 10^{-5}$. Thus, a sample of one million molecules gives rise to an expected number of doubles well within the preferred range. Such a sample is readily obtained as follows: Assume that the 5×10^{11} mRNAs are perfectly converted into 5×10^{11} vectors with tag-cDNA conjugates as inserts and that the 5×10^{11} vectors are in a reaction solution having a volume of 100 μ l. Four 10-fold serial dilutions may be carried out by transferring 10 μ l from the original solution into a vessel containing 90 μ l of an appropriate buffer, such as TE. This process may be repeated for three additional dilutions to obtain a 100 μ l solution containing 5×10^5 vector molecules per μ l. A 2 μ l aliquot from this solution yields 10^6 vectors containing tag-cDNA conjugates as inserts. This sample is then amplified by straight forward transformation of a competent host cell followed by culturing.

Of course, as mentioned above, no step in the above process proceeds with perfect efficiency. In particular, when vectors are employed to amplify a sample of tag-polynucleotide conjugates, the step of transforming a host is very inefficient. Usually, no more than 1% of the vectors are taken up by the host and replicated. Thus, for such a method of amplification, even fewer dilutions would be required to obtain a sample of 10^6 conjugates.

A repertoire of oligonucleotide tags can be conjugated to a population of polynucleotides in a number of ways, including direct enzymatic ligation, amplification, e.g. via PCR, using primers containing the tag sequences, and the like. The initial ligating step produces a very large population of tag-polynucleotide conjugates such that a single tag is generally attached to many different polynucleotides. However, as noted above, by taking a sufficiently small sample of the conjugates, the probability of obtaining "doubles," i.e. the same tag on two different polynucleotides, can be made negligible. Generally, the larger the sample the greater the probability of obtaining a double. Thus, a design trade-off exists between selecting a large sample of tag-polynucleotide conjugates--which, for example, ensures adequate coverage of a target polynucleotide in a shotgun sequencing operation or adequate representation of a rapidly changing mRNA pool, and selecting a small sample which ensures that a minimal number of doubles will be present. In most embodiments, the presence of doubles merely adds an additional source of noise or, in the case of sequencing, a minor complication in scanning and signal processing, as microparticles giving multiple fluorescent signals can simply be ignored.

As used herein, the term "substantially all" in reference to attaching tags to molecules, especially polynucleotides, is meant to reflect the statistical nature of the sampling procedure employed to obtain a population of tag-molecule conjugates essentially free of doubles. The meaning of substantially all in terms of actual

percentages of tag-molecule conjugates depends on how the tags are being employed. Preferably, for nucleic acid sequencing, substantially all means that at least eighty percent of the polynucleotides have unique tags attached. More preferably, it means that at least ninety percent of the polynucleotides have unique tags attached. Still
 5 more preferably, it means that at least ninety-five percent of the polynucleotides have unique tags attached. And, most preferably, it means that at least ninety-nine percent of the polynucleotides have unique tags attached.

Preferably, when the population of polynucleotides consists of messenger RNA (mRNA), oligonucleotides tags may be attached by reverse transcribing the
 10 mRNA with a set of primers preferably containing complements of tag sequences. An exemplary set of such primers could have the following sequence (SEQ ID NO: 1):

5' -mRNA- [A]_n -3'
 15 [T]₁₉GG[W,W,W,C]₉ACCAGCTGATC-5' -biotin

where "[W,W,W,C]₉" represents the sequence of an oligonucleotide tag of nine subunits of four nucleotides each and "[W,W,W,C]" represents the subunit sequences
 20 listed above, i.e. "W" represents T or A. The underlined sequences identify an optional restriction endonuclease site that can be used to release the polynucleotide from attachment to a solid phase support via the biotin, if one is employed. For the above primer, the complement attached to a microparticle could have the form:

25 5' - [G,W,W,W]₉TGG-linker-microparticle

After reverse transcription, the mRNA is removed, e.g. by RNase H digestion, and the second strand of the cDNA is synthesized using, for example, a primer of the following form (SEQ ID NO: 2):

30 5' -NRRGATCYNNN-3'

where N is any one of A, T, G, or C; R is a purine-containing nucleotide, and Y is a pyrimidine-containing nucleotide. This particular primer creates a Bst Y1 restriction
 35 site in the resulting double stranded DNA which, together with the Sal I site, facilitates cloning into a vector with, for example, Bam HI and Xho I sites. After Bst Y1 and Sal I digestion, the exemplary conjugate would have the form:

5'-RCGACCA[C,W,W,W]₉GG[T]₁₉- cDNA -NNNR
 GGT[G,W,W,W]₉CC[A]₁₉- rDNA -NNNYCTAG-5'

The polynucleotide-tag conjugates may then be manipulated using standard molecular biology techniques. For example, the above conjugate--which is actually a mixture--may be inserted into commercially available cloning vectors, e.g. Stratagene Cloning System (La Jolla, CA); transfected into a host, such as a commercially available host bacteria; which is then cultured to increase the number of conjugates. The cloning vectors may then be isolated using standard techniques, e.g. Sambrook et al, Molecular Cloning, Second Edition (Cold Spring Harbor Laboratory, New York, 1989). Alternatively, appropriate adaptors and primers may be employed so that the conjugate population can be increased by PCR.

Preferably, when the ligase-based method of sequencing is employed, the Bst Y1 and Sal I digested fragments are cloned into a Bam HI-/Xho I-digested vector having the following single-copy restriction sites (SEQ ID NO: 3):

5'-GAGGATGCCTTTATGGATCCACTCGAGATCCCAATCCA-3'
 FokI BamHI XhoI

This adds the Fok I site which will allow initiation of the sequencing process discussed more fully below.

Tags can be conjugated to cDNAs of existing libraries by standard cloning methods. cDNAs are excised from their existing vector, isolated, and then ligated into a vector containing a repertoire of tags. Preferably, the tag-containing vector is linearized by cleaving with two restriction enzymes so that the excised cDNAs can be ligated in a predetermined orientation. The concentration of the linearized tag-containing vector is in substantial excess over that of the cDNA inserts so that ligation provides an inherent sampling of tags.

A general method for exposing the single stranded tag after amplification involves digesting a target polynucleotide-containing conjugate with the 5'→3' exonuclease activity of T4 DNA polymerase, or a like enzyme. When used in the presence of a single deoxynucleoside triphosphate, such a polymerase will cleave nucleotides from 3' recessed ends present on the non-template strand of a double stranded fragment until a complement of the single deoxynucleoside triphosphate is reached on the template strand. When such a nucleotide is reached the 5'→3' digestion effectively ceases, as the polymerase's extension activity adds nucleotides at a higher rate than the excision activity removes nucleotides. Consequently, single

stranded tags constructed with three nucleotides are readily prepared for loading onto solid phase supports.

The technique may also be used to preferentially methylate interior Fok I sites of a target polynucleotide while leaving a single Fok I site at the terminus of the polynucleotide unmethylated. First, the terminal Fok I site is rendered single stranded using a polymerase with deoxycytidine triphosphate. The double stranded portion of the fragment is then methylated, after which the single stranded terminus is filled in with a DNA polymerase in the presence of all four nucleoside triphosphates, thereby regenerating the Fok I site. Clearly, this procedure can be generalized to endonucleases other than Fok I.

After the oligonucleotide tags are prepared for specific hybridization, e.g. by rendering them single stranded as described above, the polynucleotides are mixed with microparticles containing the complementary sequences of the tags under conditions that favor the formation of perfectly matched duplexes between the tags and their complements. There is extensive guidance in the literature for creating these conditions. Exemplary references providing such guidance include Wetmur, *Critical Reviews in Biochemistry and Molecular Biology*, 26: 227-259 (1991); Sambrook et al, *Molecular Cloning: A Laboratory Manual*, 2nd Edition (Cold Spring Harbor Laboratory, New York, 1989); and the like. Preferably, the hybridization conditions are sufficiently stringent so that only perfectly matched sequences form stable duplexes. Under such conditions the polynucleotides specifically hybridized through their tags may be ligated to the complementary sequences attached to the microparticles. Finally, the microparticles are washed to remove polynucleotides with unligated and/or mismatched tags.

When CPG microparticles conventionally employed as synthesis supports are used, the density of tag complements on the microparticle surface is typically greater than that necessary for some sequencing operations. That is, in sequencing approaches that require successive treatment of the attached polynucleotides with a variety of enzymes, densely spaced polynucleotides may tend to inhibit access of the relatively bulky enzymes to the polynucleotides. In such cases, the polynucleotides are preferably mixed with the microparticles so that tag complements are present in significant excess, e.g. from 10:1 to 100:1, or greater, over the polynucleotides. This ensures that the density of polynucleotides on the microparticle surface will not be so high as to inhibit enzyme access. Preferably, the average inter-polynucleotide spacing on the microparticle surface is on the order of 30-100 nm. Guidance in selecting ratios for standard CPG supports and Ballotini beads (a type of solid glass support) is found in Maskos and Southern, *Nucleic Acids Research*, 20: 1679-1684 (1992). Preferably, for sequencing applications, standard CPG beads of diameter in the range

of 20-50 μm are loaded with about 10^5 polynucleotides, and GMA beads of diameter in the range of 5-10 μm are loaded with a few tens of thousand of polynucleotides, e.g. 4×10^4 to 6×10^4 .

In the preferred embodiment, tag complements are synthesized on
 5 microparticles combinatorially; thus, at the end of the synthesis, one obtains a complex mixture of microparticles from which a sample is taken for loading tagged polynucleotides. The size of the sample of microparticles will depend on several factors, including the size of the repertoire of tag complements, the nature of the apparatus for used for observing loaded microparticles--e.g. its capacity, the tolerance
 10 for multiple copies of microparticles with the same tag complement (i.e. "bead doubles"), and the like. The following table provide guidance regarding microparticle sample size, microparticle diameter, and the approximate physical dimensions of a packed array of microparticles of various diameters.

15

Microparticle diameter	5 μm	10 μm	20 μm	40 μm
Max. no. polynucleotides loaded at 1 per 10^5 sq. angstrom		3×10^5	1.26×10^6	5×10^6
Approx. area of monolayer of 10^6 microparticles	.45 x .45 cm	1 x 1 cm	2 x 2 cm	4 x 4 cm

20 The probability that the sample of microparticles contains a given tag complement or is present in multiple copies is described by the Poisson distribution, as indicated in the following table.

25

Table VII

Number of microparticles in sample (as fraction of repertoire size), m	Fraction of repertoire of tag complements present in sample, $1-e^{-m}$	Fraction of microparticles in sample with unique tag complement attached, $m(e^{-m})/2$	Fraction of microparticles in sample carrying same tag complement as one other microparticle in sample ("bead doubles"), $m^2(e^{-m})/2$
1.000	0.63	0.37	0.18
.693	0.50	0.35	0.12
.405	0.33	0.27	0.05
.285	0.25	0.21	0.03
.223	0.20	0.18	0.02
.105	0.10	0.09	0.005
.010	0.01	0.01	

High Specificity Sorting and Panning

5 The kinetics of sorting depends on the rate of hybridization of oligonucleotide tags to their tag complements which, in turn, depends on the complexity of the tags in the hybridization reaction. Thus, a trade off exists between sorting rate and tag complexity, such that an increase in sorting rate may be achieved at the cost of reducing the complexity of the tags involved in the hybridization reaction. As explained below, the effects of this trade off may be ameliorated by "panning."

10 Specificity of the hybridizations may be increased by taking a sufficiently small sample so that both a high percentage of tags in the sample are unique and the nearest neighbors of substantially all the tags in a sample differ by at least two words. This latter condition may be met by taking a sample that contains a number of tag-polynucleotide conjugates that is about 0.1 percent or less of the size of the repertoire being employed. For example, if tags are constructed with eight words selected from Table II, a repertoire of 8^8 , or about 1.67×10^7 , tags and tag complements are produced. In a library of tag-cDNA conjugates as described above, a 0.1 percent sample means that about 16,700 different tags are present. If this were loaded directly onto a repertoire-equivalent of microparticles, or in this example a sample of 1.67×10^7 microparticles, then only a sparse subset of the sampled microparticles would be loaded. The density of loaded microparticles can be increase--for example, for more efficient sequencing--by undertaking a "panning" step in which the sampled tag-cDNA conjugates are used to separate loaded microparticles from unloaded microparticles. Thus, in the example above, even though a "0.1 percent" sample

contains only 16,700 cDNAs, the sampling and panning steps may be repeated until as many loaded microparticles as desired are accumulated.

A panning step may be implemented by providing a sample of tag-cDNA conjugates each of which contains a capture moiety at an end opposite, or distal to, the oligonucleotide tag. Preferably, the capture moiety is of a type which can be released from the tag-cDNA conjugates, so that the tag-cDNA conjugates can be sequenced with a single-base sequencing method. Such moieties may comprise biotin, digoxigenin, or like ligands, a triplex binding region, or the like. Preferably, such a capture moiety comprises a biotin component. Biotin may be attached to tag-cDNA conjugates by a number of standard techniques. If appropriate adapters containing PCR primer binding sites are attached to tag-cDNA conjugates, biotin may be attached by using a biotinylated primer in an amplification after sampling. Alternatively, if the tag-cDNA conjugates are inserts of cloning vectors, biotin may be attached after excising the tag-cDNA conjugates by digestion with an appropriate restriction enzyme followed by isolation and filling in a protruding strand distal to the tags with a DNA polymerase in the presence of biotinylated uridine triphosphate.

After a tag-cDNA conjugate is captured, it may be released from the biotin moiety in a number of ways, such as by a chemical linkage that is cleaved by reduction, e.g. Herman et al, Anal. Biochem., 156: 48-55 (1986), or that is cleaved photochemically, e.g. Olejnik et al, Nucleic Acids Research, 24: 361-366 (1996), or that is cleaved enzymatically by introducing a restriction site in the PCR primer. The latter embodiment can be exemplified by considering the library of tag-polynucleotide conjugates described above:

5'-RCGACCA[C,W,W,W]₉GG[T]₁₉- cDNA -NNNR
GGT[G,W,W,W]₉CC[A]₁₉- rDNA -NNNYCTAG-5'

The following adapters may be ligated to the ends of these fragments to permit amplification by PCR:

5' - XXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXYGAT

Right Adapter

GATCZZACTAGTZZZZZZZZZZZZ-3'
ZZTGATCAZZZZZZZZZZZZ

Left Adapter

ZZTGATCAZZZZZZZZZZZZ-5'-biotin

5

Left Primer

where "ACTAGT" is a Spe I recognition site (which leaves a staggered cleavage ready for single base sequencing), and the X's and Z's are nucleotides selected so that the annealing and dissociation temperatures of the respective primers are approximately the same. After ligation of the adapters and amplification by PCR using the biotinylated primer, the tags of the conjugates are rendered single stranded by the exonuclease activity of T4 DNA polymerase and conjugates are combined with a sample of microparticles, e.g. a repertoire equivalent, with tag complements attached. After annealing under stringent conditions (to minimize mis-attachment of tags), the conjugates are preferably ligated to their tag complements and the loaded microparticles are separated from the unloaded microparticles by capture with avidinated magnetic beads, or like capture technique.

Returning to the example, this process results in the accumulation of about 10,500 ($=16,700 \times .63$) loaded microparticles with different tags, which may be released from the magnetic beads by cleavage with Spe I. By repeating this process 40-50 times with new samples of microparticles and tag-cDNA conjugates, $4-5 \times 10^5$ cDNAs can be accumulated by pooling the released microparticles. The pooled microparticles may then be simultaneously sequenced by a single-base sequencing technique.

Determining how many times to repeat the sampling and panning steps--or more generally, determining how many cDNAs to analyze, depends on one's objective. If the objective is to monitor the changes in abundance of relatively common sequences, e.g. making up 5% or more of a population, then relatively small samples, i.e. a small fraction of the total population size, may allow statistically significant estimates of relative abundances. On the other hand, if one seeks to monitor the abundances of rare sequences, e.g. making up 0.1% or less of a population, then large samples are required. Generally, there is a direct relationship between sample size and the reliability of the estimates of relative abundances based on the sample. There is extensive guidance in the literature on determining appropriate sample sizes for making reliable statistical estimates, e.g. Koller et al, Nucleic Acids Research, 23:185-191 (1994); Good, Biometrika, 40: 16-264 (1953); Bunge et al, J. Am. Stat. Assoc., 88: 364-373 (1993); and the like. Preferably, for

monitoring changes in gene expression based on the analysis of a series of cDNA libraries containing 10^5 to 10^8 independent clones of 3.0 - 3.5×10^4 different sequences, a sample of at least 10^4 sequences are accumulated for analysis of each library. More preferably, a sample of at least 10^5 sequences are accumulated for the analysis of each library; and most preferably, a sample of at least 5×10^5 sequences are accumulated for the analysis of each library. Alternatively, the number of sequences sampled is preferably sufficient to estimate the relative abundance of a sequence present at a frequency within the range of 0.1% to 5% with a 95% confidence limit no larger than 0.1% of the population size.

10

Single Base DNA Sequencing

The present invention can be employed with conventional methods of DNA sequencing, e.g. as disclosed by Hultman et al, Nucleic Acids Research, 17: 4937-4946 (1989). However, for parallel, or simultaneous, sequencing of multiple polynucleotides, a DNA sequencing methodology is preferred that requires neither electrophoretic separation of closely sized DNA fragments nor analysis of cleaved nucleotides by a separate analytical procedure, as in peptide sequencing. Preferably, the methodology permits the stepwise identification of nucleotides, usually one at a time, in a sequence through successive cycles of treatment and detection. Such methodologies are referred to herein as "single base" sequencing methods. Single base approaches are disclosed in the following references: Cheeseman, U.S. patent 5,302,509; Tsien et al, International application WO 91/06678; Rosenthal et al, International application WO 93/21340; Canard et al, Gene, 148: 1-6 (1994); and Metzker et al, Nucleic Acids Research, 22: 4259-4267 (1994).

A "single base" method of DNA sequencing which is suitable for use with the present invention and which requires no electrophoretic separation of DNA fragments is described in International application PCT/US95/03678. Briefly, the method comprises the following steps: (a) ligating a probe to an end of the polynucleotide having a protruding strand to form a ligated complex, the probe having a complementary protruding strand to that of the polynucleotide and the probe having a nuclease recognition site; (b) removing unligated probe from the ligated complex; (c) identifying one or more nucleotides in the protruding strand of the polynucleotide by the identity of the ligated probe; (d) cleaving the ligated complex with a nuclease; and (e) repeating steps (a) through (d) until the nucleotide sequence of the polynucleotide, or a portion thereof, is determined.

A single signal generating moiety, such as a single fluorescent dye, may be employed when sequencing several different target polynucleotides attached to different spatially addressable solid phase supports, such as fixed microparticles, in a

parallel sequencing operation. This may be accomplished by providing four sets of probes that are applied sequentially to the plurality of target polynucleotides on the different microparticles. An exemplary set of such probes are shown below:

5

Set 1	Set 2	Set 3	Set 4
ANNNN...NN N...NNTT...T*	dANNNN...NN d N...NNTT...T	dANNNN...NN N...NNTT...T	dANNNN...NN N...NNTT...T
dCNNNN...NN N...NNTT...T	CNNNN...NN N...NNTT...T*	dCNNNN...NN N...NNTT...T	dCNNNN...NN N...NNTT...T
dGNNNN...NN N...NNTT...T	dGNNNN...NN N...NNTT...T	GNNNN...NN N...NNTT...T*	dGNNNN...NN N...NNTT...T
dTNNNN...NN N...NNTT...T	dTNNNN...NN N...NNTT...T	dTNNNN...NN N...NNTT...T	TNNNN...NN N...NNTT...T*

where each of the listed probes represents a mixture of $4^3=64$ oligonucleotides such that the identity of the 3' terminal nucleotide of the top strand is fixed and the other positions in the protruding strand are filled by every 3-mer permutation of nucleotides, or complexity reducing analogs. The listed probes are also shown with a single stranded poly-T tail with a signal generating moiety attached to the terminal thymidine, shown as "T*". The "d" on the unlabeled probes designates a ligation-blocking moiety or absence of 3'-hydroxyl, which prevents unlabeled probes from being ligated. Preferably, such 3'-terminal nucleotides are dideoxynucleotides. In this embodiment, the probes of set 1 are first applied to the plurality of target polynucleotides and treated with a ligase so that target polynucleotides having a thymidine complementary to the 3' terminal adenosine of the labeled probes are ligated. The unlabeled probes are simultaneously applied to minimize inappropriate ligations. The locations of the target polynucleotides that form ligated complexes with probes terminating in "A" are identified by the signal generated by the label carried on the probe. After washing and cleavage, the probes of set 2 are applied. In this case, target polynucleotides forming ligated complexes with probes terminating in "C" are identified by location. Similarly, the probes of sets 3 and 4 are applied and locations of positive signals identified. This process of sequentially applying the four sets of probes continues until the desired number of nucleotides are identified on the target polynucleotides. Clearly, one of ordinary skill could construct similar sets of probes that could have many variations, such as having protruding strands of different lengths, different moieties to block ligation of unlabeled probes, different means for labeling probes, and the like.

Apparatus for Sequencing Populations of Polynucleotides

An objective of the invention is to sort identical molecules, particularly polynucleotides, onto the surfaces of microparticles by the specific hybridization of tags and their complements. Once such sorting has taken place, the presence of the molecules or operations performed on them can be detected in a number of ways depending on the nature of the tagged molecule, whether microparticles are detected separately or in "batches," whether repeated measurements are desired, and the like. Typically, the sorted molecules are exposed to ligands for binding, e.g. in drug development, or are subjected chemical or enzymatic processes, e.g. in polynucleotide sequencing. In both of these uses it is often desirable to simultaneously observe signals corresponding to such events or processes on large numbers of microparticles. Microparticles carrying sorted molecules (referred to herein as "loaded" microparticles) lend themselves to such large scale parallel operations, e.g. as demonstrated by Lam et al (cited above).

Preferably, whenever light-generating signals, e.g. chemiluminescent, fluorescent, or the like, are employed to detect events or processes, loaded microparticles are spread on a planar substrate, e.g. a glass slide, for examination with a scanning system, such as described in International patent applications PCT/US91/09217, PCT/NL90/00081, and PCT/US95/01886. The scanning system should be able to reproducibly scan the substrate and to define the positions of each microparticle in a predetermined region by way of a coordinate system. In polynucleotide sequencing applications, it is important that the positional identification of microparticles be repeatable in successive scan steps.

Such scanning systems may be constructed from commercially available components, e.g. x-y translation table controlled by a digital computer used with a detection system comprising one or more photomultiplier tubes, or alternatively, a CCD array, and appropriate optics, e.g. for exciting, collecting, and sorting fluorescent signals. In some embodiments a confocal optical system may be desirable. An exemplary scanning system suitable for use in four-color sequencing is illustrated diagrammatically in Figure 5. Substrate 300, e.g. a microscope slide with fixed microparticles, is placed on x-y translation table 302, which is connected to and controlled by an appropriately programmed digital computer 304 which may be any of a variety of commercially available personal computers, e.g. 486-based machines or PowerPC model 7100 or 8100 available from Apple Computer (Cupertino, CA). Computer software for table translation and data collection functions can be provided by commercially available laboratory software, such as Lab Windows, available from National Instruments.

Substrate 300 and table 302 are operationally associated with microscope 306 having one or more objective lenses 308 which are capable of collecting and delivering light to microparticles fixed to substrate 300. Excitation beam 310 from light source 312, which is preferably a laser, is directed to beam splitter 314, e.g. a dichroic mirror, which re-directs the beam through microscope 306 and objective lens 308 which, in turn, focuses the beam onto substrate 300. Lens 308 collects fluorescence 316 emitted from the microparticles and directs it through beam splitter 314 to signal distribution optics 318 which, in turn, directs fluorescence to one or more suitable opto-electronic devices for converting some fluorescence characteristic, e.g. intensity, lifetime, or the like, to an electrical signal. Signal distribution optics 318 may comprise a variety of components standard in the art, such as bandpass filters, fiber optics, rotating mirrors, fixed position mirrors and lenses, diffraction gratings, and the like. As illustrated in Figure 2, signal distribution optics 318 directs fluorescence 316 to four separate photomultiplier tubes, 330, 332, 334, and 336, whose output is then directed to pre-amps and photon counters 350, 352, 354, and 356. The output of the photon counters is collected by computer 304, where it can be stored, analyzed, and viewed on video 360. Alternatively, signal distribution optics 318 could be a diffraction grating which directs fluorescent signal 318 onto a CCD array.

The stability and reproducibility of the positional localization in scanning will determine, to a large extent, the resolution for separating closely spaced microparticles. Preferably, the scanning systems should be capable of resolving closely spaced microparticles, e.g. separated by a particle diameter or less. Thus, for most applications, e.g. using CPG microparticles, the scanning system should at least have the capability of resolving objects on the order of 10-100 μm . Even higher resolution may be desirable in some embodiments, but with increase resolution, the time required to fully scan a substrate will increase; thus, in some embodiments a compromise may have to be made between speed and resolution. Increases in scanning time can be achieved by a system which only scans positions where microparticles are known to be located, e.g. from an initial full scan. Preferably, microparticle size and scanning system resolution are selected to permit resolution of fluorescently labeled microparticles randomly disposed on a plane at a density between about ten thousand to one hundred thousand microparticles per cm^2 .

In sequencing applications, loaded microparticles can be fixed to the surface of a substrate in variety of ways. The fixation should be strong enough to allow the microparticles to undergo successive cycles of reagent exposure and washing without significant loss. When the substrate is glass, its surface may be derivatized with an alkylamino linker using commercially available reagents, e.g. Pierce Chemical, which

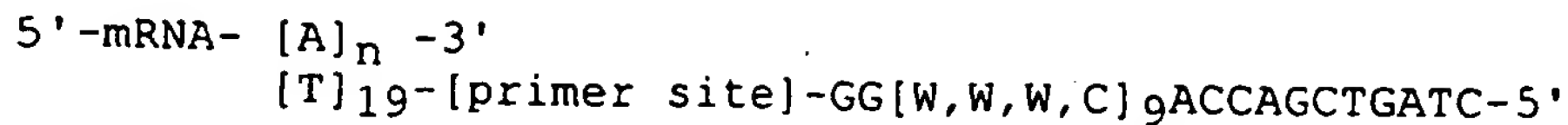
in turn may be cross-linked to avidin, again using conventional chemistries, to form an avidinated surface. Biotin moieties can be introduced to the loaded microparticles in a number of ways. For example, a fraction, e.g. 10-15 percent, of the cloning vectors used to attach tags to polynucleotides are engineered to contain a unique
5 restriction site (providing sticky ends on digestion) immediately adjacent to the polynucleotide insert at an end of the polynucleotide opposite of the tag. The site is excised with the polynucleotide and tag for loading onto microparticles. After loading, about 10-15 percent of the loaded polynucleotides will possess the unique restriction site distal from the microparticle surface. After digestion with the
10 associated restriction endonuclease, an appropriate double stranded adaptor containing a biotin moiety is ligated to the sticky end. The resulting microparticles are then spread on the avidinated glass surface where they become fixed via the biotin-avidin linkages.

Alternatively and preferably when sequencing by ligation is employed, in the
15 initial ligation step a mixture of probes is applied to the loaded microparticle: a fraction of the probes contain a type II's restriction recognition site, as required by the sequencing method, and a fraction of the probes have no such recognition site, but instead contain a biotin moiety at its non-ligating end. Preferably, the mixture comprises about 10-15 percent of the biotinylated probe.

20 In still another alternative, when DNA-loaded microparticles are applied to a glass substrate, the DNA may nonspecifically adsorb to the glass surface upon several hours, e.g. 24 hours, incubation to create a bond sufficiently strong to permit repeated exposures to reagents and washes without significant loss of microparticles. Preferably, such a glass substrate is a flow cell, which may comprise a channel etched
25 in a glass slide. Preferably, such a channel is closed so that fluids may be pumped through it and has a depth sufficiently close to the diameter of the microparticles so that a monolayer of microparticles is trapped within a defined observation region.

Identification of Novel Polynucleotides in cDNA Libraries

30 Novel polynucleotides in a cDNA library can be identified by constructing a library of cDNA molecules attached to microparticles, as described above. A large fraction of the library, or even the entire library, can then be partially sequenced in parallel. After isolation of mRNA, and perhaps normalization of the population as taught by Soares et al, Proc. Natl. Acad. Sci., 91: 9228-9232 (1994), or like
35 references, the following primer may be hybridized to the polyA tails for first strand synthesis with a reverse transcriptase using conventional protocols (SEQ ID NO: 1):



where [W,W,W,C]₉ represents a tag as described above, "ACCAGCTGATC" is an optional sequence forming a restriction site in double stranded form, and "primer site" is a sequence common to all members of the library that is later used as a primer binding site for amplifying polynucleotides of interest by PCR.

After reverse transcription and second strand synthesis by conventional techniques, the double stranded fragments are inserted into a cloning vector as described above and amplified. The amplified library is then sampled and the sample amplified. The cloning vectors from the amplified sample are isolated, and the tagged cDNA fragments excised and purified. After rendering the tag single stranded with a polymerase as described above, the fragments are methylated and sorted onto microparticles in accordance with the invention. Preferably, as described above, the cloning vector is constructed so that the tagged cDNAs can be excised with an endonuclease, such as Fok I, that will allow immediate sequencing by the preferred single base method after sorting and ligation to microparticles.

Stepwise sequencing is then carried out simultaneously on the whole library, or one or more large fractions of the library, in accordance with the invention until a sufficient number of nucleotides are identified on each cDNA for unique representation in the genome of the organism from which the library is derived. For example, if the library is derived from mammalian mRNA then a randomly selected sequence 14-15 nucleotides long is expected to have unique representation among the 2-3 thousand megabases of the typical mammalian genome. Of course identification of far fewer nucleotides would be sufficient for unique representation in a library derived from bacteria, or other lower organisms. Preferably, at least 20-30 nucleotides are identified to ensure unique representation and to permit construction of a suitable primer as described below. The tabulated sequences may then be compared to known sequences to identify unique cDNAs.

Unique cDNAs are then isolated by conventional techniques, e.g. constructing a probe from the PCR amplicon produced with primers directed to the prime site and the portion of the cDNA whose sequence was determined. The probe may then be used to identify the cDNA in a library using a conventional screening protocol.

The above method for identifying new cDNAs may also be used to fingerprint mRNA populations, either in isolated measurements or in the context of a dynamically changing population. Partial sequence information is obtained simultaneously from a large sample, e.g. ten to a hundred thousand, or more, of cDNAs attached to separate microparticles as described in the above method.

Example 1**Construction of a Tag Library**

An exemplary tag library is constructed as follows to form the chemically
 5 synthesized 9-word tags of nucleotides A, G, and T defined by the formula:



where "[$\text{}^4\text{(A,G,T)}_9$]" indicates a tag mixture where each tag consists of nine 4-mer
 10 words of A, G, and T; and "p" indicate a 5' phosphate. This mixture is ligated to the
 following right and left primer binding regions (SEQ ID NO: 4 and SEQ ID NO 5):

5' - AGTGGCTGGGCATCGGACCG
 TCACCGACCCGTAGCCp

5' - GGGGCCCAGTCAGCGTCGAT
 GGGTCAGTCGCAGCTA

15

LEFT

RIGHT

The right and left primer binding regions are ligated to the above tag mixture, after
 which the single stranded portion of the ligated structure is filled with DNA
 20 polymerase then mixed with the right and left primers indicated below and amplified
 to give a tag library (SEQ ID NO: 6).

Left Primer

25

5' - AGTGGCTGGGCATCGGACCG

5' - AGTGGCTGGGCATCGGACCG- [$\text{}^4\text{(A,G,T)}_9$]-GGGGCCCAGTCAGCGTCGAT
 TCACCGACCCGTAGCCTGGC- [$\text{}^4\text{(A,G,T)}_9$]-CCCCGGGTCAGTCGCAGCTA

30

CCCCGGGTCAGTCGCAGCTA-5'

Right Primer

35 The underlined portion of the left primer binding region indicates a Rsr II recognition
 site. The left-most underlined region of the right primer binding region indicates
 recognition sites for Bsp 120I, Apa I, and Eco O 109I, and a cleavage site for Hga I.
 The right-most underlined region of the right primer binding region indicates the
 recognition site for Hga I. Optionally, the right or left primers may be synthesized
 40 with a biotin attached (using conventional reagents, e.g. available from Clontech
 Laboratories, Palo Alto, CA) to facilitate purification after amplification and/or
 cleavage.

NOT FURNISHED UPON FILING



20

25

30

30

35

Japan); monoclonal mouse anti-rat CYP1A1, monoclonal mouse anti-rat CYP2C11, goat anti-rat CYP2E1, and monoclonal mouse anti-rat CYP2B1 from Oxford Biochemical Research, Inc. (Oxford, MI). Secondary antibodies (goat anti-rabbit IgG, rabbit anti-goat IgG and goat anti-mouse IgG) are available from Jackson
5 ImmunoResearch Laboratories (West Grove, PA).

Animals are administered either PB (100 mg/kg), BNF (100 mg/kg), MET (100 mg/kg), DEX (100 mg/kg), or CLO (250 mg/kg) for 4 consecutive days via intraperitoneal injection following a dosing regimen similar to that described by Wang et al, Arch. Biochem. Biophys. 290: 355-361 (1991). Animals treated with
10 H₂O and CO are used as controls. Two hours following the last injection (day 4), animals are killed, and the livers are removed. Livers are immediately frozen and stored at -70°C.

Total RNA is prepared from frozen liver tissue using a modification of the method described by Xie et al, Biotechniques, 11: 326-327 (1991). Approximately
15 100-200 mg of liver tissue is homogenized in the RNA extraction buffer described by Xie et al to isolate total RNA. The resulting RNA is reconstituted in diethylpyrocarbonate-treated water, quantified spectrophotometrically at 260 nm, and adjusted to a concentration of 100 µg/ml. Total RNA is stored in
- diethylpyrocarbonate-treated water for up to 1 year at -70°C without any apparent
20 degradation. RT-PCR and sequencing are performed on samples from these preparations.

For sequencing, samples of RNA corresponding to about 0.5 µg of poly(A)⁺ RNA are used to construct libraries of tag-cDNA conjugates following the protocol described in the section entitled "Attaching Tags to Polynucleotides for Sorting onto
25 Solid Phase Supports," with the following exception: the tag repertoire is constructed from six 4-nucleotide words from Table II. Thus, the complexity of the repertoire is 8⁶ or about 2.6 x 10⁵. For each tag-cDNA conjugate library constructed, ten samples of about ten thousand clones are taken for amplification and sorting. Each of the amplified samples is separately applied to a fixed monolayer of about 10⁶ 10 µm
30 diameter GMA beads containing tag complements. That is, the "sample" of tag complements in the GMA bead population on each monolayer is about four fold the total size of the repertoire, thus ensuring there is a high probability that each of the sampled tag-cDNA conjugates will find its tag complement on the monolayer. After the oligonucleotide tags of the amplified samples are rendered single stranded as
35 described above, the tag-cDNA conjugates of the samples are separately applied to the monolayers under conditions that permit specific hybridization only between oligonucleotide tags and tag complements forming perfectly matched duplexes. Concentrations of the amplified samples and hybridization times are selected to

permit the loading of about 5×10^4 to 2×10^5 tag-cDNA conjugates on each bead where perfect matches occur. After ligation, 9-12 nucleotide portions of the attached cDNAs are determined in parallel by the single base sequencing technique described by Brenner in International patent application PCT/US95/03678. Frequency
5 distributions for the gene expression profiles are assembled from the sequence information obtained from each of the ten samples.

RT-PCRs of selected mRNAs corresponding to cytochrome P-450 genes and the constitutively expressed cyclophilin gene are carried out as described in Morris et al (cited above). Briefly, a 20 μ L reaction mixture is prepared containing 1x reverse
10 transcriptase buffer (Gibco BRL), 10 nM dithiothreitol, 0.5 nM dNTPs, 2.5 μ M oligo d(T)₁₅ primer, 40 units RNasin (Promega, Madison, WI), 200 units RNase H-reverse transcriptase (Gibco BRL), and 400 ng of total RNA (in diethylpyrocarbonate-treated water). The reaction is incubated for 1 hour at 37°C followed by inactivation of the enzyme at 95°C for 5 min. The resulting cDNA is stored at -20°C until used. For
15 PCR amplification of cDNA, a 10 μ L reaction mixture is prepared containing 10x polymerase reaction buffer, 2 mM MgCl₂, 1 unit Taq DNA polymerase (Perkin-Elmer, Norwalk, CT), 20 ng cDNA, and 200 nM concentration of the 5' and 3' specific PCR primers of the sequences described in Morris et al (cited above). PCRs
-are carried out in a Perkin-Elmer 9600 thermal cycler for 23 cycles using melting,
20 annealing, and extension conditions of 94°C for 30 sec., 56°C for 1 min., and 72°C for 1 min., respectively. Amplified cDNA products are separated by PAGE using 5% native gels. Bands are detected by staining with ethidium bromide.

Western blots of the liver proteins are carried out using standard protocols after separation by SDS-PAGE. Briefly, proteins are separated on 10% SDS-PAGE
25 gels under reducing conditions and immunoblotted for detection of P-450 isoenzymes using a modification of the methods described in Harris et al, Proc. Natl. Acad. Sci., 88: 1407-1410 (1991). Protein are loaded at 50 μ g/lane and resolved under constant current (250 V) for approximately 4 hours at 2°C. Proteins are transferred to
nitrocellulose membranes (Bio-Rad, Hercules, CA) in 15 mM Tris buffer containing
30 120 mM glycine and 20% (v/v) methanol. The nitrocellulose membranes are blocked with 2.5% BSA and immunoblotted for P-450 isoenzymes using primary monoclonal and polyclonal antibodies and secondary alkaline phosphatase conjugated anti-IgG. Immunoblots are developed with the Bio-Rad alkaline phosphatase substrate kit.

The three types of measurements of P-450 isoenzyme induction showed
35 substantial agreement.

APPENDIX Ia
Exemplary computer program for generating
minimally cross hybridizing sets
(single stranded tag/single stranded tag complement)

```

Program minxh
c
c
c
      integer*2 sub1(6),mset1(1000,6),mset2(1000,6)
      dimension nbase(6)
c
c
      write(*,*) 'ENTER SUBUNIT LENGTH'
      read(*,100) nsub
100    format(i1)
      open(1,file='sub4.dat',form='formatted',status='new')
c
c
      nset=0
      do 7000 m1=1,3
        do 7000 m2=1,3
          do 7000 m3=1,3
            do 7000 m4=1,3
              sub1(1)=m1
              sub1(2)=m2
              sub1(3)=m3
              sub1(4)=m4
c
c
      ndiff=3
c
c
c      Generate set of subunits differing from
c      sub1 by at least ndiff nucleotides.
c      Save in mset1.
c
c
      jj=1
      do 900 j=1,nsub
900    mset1(1,j)=sub1(j)
c
c
      do 1000 k1=1,3
        do 1000 k2=1,3
          do 1000 k3=1,3
            do 1000 k4=1,3
c
c
              nbase(1)=k1
              nbase(2)=k2
              nbase(3)=k3
              nbase(4)=k4

```

```

c
      n=0
      do 1200 j=1, nsub
        if (sub1(j).eq.1 .and. nbase(j).ne.1 .or.
1         sub1(j).eq.2 .and. nbase(j).ne.2 .or.
3         sub1(j).eq.3 .and. nbase(j).ne.3) then
          n=n+1
        endif
1200      continue
c
c
      if (n.ge.ndiff) then
c
c
c          If number of mismatches
c          is greater than or equal
c          to ndiff then record
c          subunit in matrix mset
c
c
c          jj=jj+1
c          do 1100 i=1, nsub
1100      mset1(jj,i)=nbase(i)
c          endif
c
c
c      1000 continue
c
c
c          do 1325 j2=1, nsub
c          mset2(1,j2)=mset1(1,j2)
1325      mset2(2,j2)=mset1(2,j2)
c
c
c          Compare subunit 2 from
c          mset1 with each successive
c          subunit in mset1, i.e. 3,
c          4,5, ... etc. Save those
c          with mismatches .ge. ndiff
c          in matrix mset2 starting at
c          position 2.
c          Next transfer contents
c          of mset2 into mset1 and
c          start
c          comparisons again this time
c          starting with subunit 3.
c          Continue until all subunits
c          undergo the comparisons.
c
c
c      npass=0
c
c
c      1700 continue
c          kk=npass+2
c          npass=npass+1
c

```

```

c
do 1500 m=npass+2,jj
  n=0
  do 1600 j=1,nsub
    if(mset1(npass+1,j).eq.1.and.mset1(m,j).ne.1.or.
2      mset1(npass+1,j).eq.2.and.mset1(m,j).ne.2.or.
2      mset1(npass+1,j).eq.3.and.mset1(m,j).ne.3) then
      n=n+1
    endif
1600    continue
    if(n.ge.ndiff) then
      kk=kk+1
      do 1625 i=1,nsub
1625        mset2(kk,i)=mset1(m,i)
      endif
1500    continue
  c
  c
  c
  c
  c
  c
  c
  c
  c
  c
  do 2000 k=1, kk
    do 2000 m=1,nsub
2000      mset1(k,m)=mset2(k,m)
    if(kk.lt.jj) then
      jj=kk
      goto 1700
    endif
  c
  c
  nset=nset+1
  write(1,7009)
7009  format(/)
  do 7008 k=1, kk
7008    write(1,7010) (mset1(k,m),m=1,nsub)
7010  format(4i1)
  write(*,*)
  write(*,120) kk,nset
120  format(1x,'Subunits in set=',i5,2x,'Set No=',i5)
7000  continue
  close(1)
  c
  c
  end
  c
  c
  *****
  *****

```

APPENDIX Ib
Exemplary computer program for generating
minimally cross hybridizing sets
(single stranded tag/single stranded tag complement)

```

Program tagN
C
C
C      Program tagN generates minimally cross-hybridizing
C      sets of subunits given i) N--subunit length, and ii)
C      an initial subunit sequence. tagN assumes that only
C      3 of the four natural nucleotides are used in the tags.
C
C
C      character*1 sub1(20)
C      integer*2 mset(10000,20), nbase(20)
C
C
C      write(*,*)'ENTER SUBUNIT LENGTH'
C      read(*,100) nsub
100    format(i2)
C
C
C      write(*,*)'ENTER SUBUNIT SEQUENCE'
C      read(*,110) (sub1(k),k=1,nsub)
110    format(20a1)
C
C
C      ndiff=10
C
C
C      Let a=1 c=2 g=3 & t=4
C
C
C      do 800 kk=1,nsub
C      if(sub1(kk).eq.'a') then
C      mset(1, kk)=1
C      endif
C      if(sub1(kk).eq.'c') then
C      mset(1, kk)=2
C      endif
C      if(sub1(kk).eq.'g') then
C      mset(1, kk)=3
C      endif
C      if(sub1(kk).eq.'t') then
C      mset(1, kk)=4
C      endif
800    continue
C
C
C      Generate set of subunits differing from
C      sub1 by at least ndiff nucleotides.
C
C
C      jj=1
C
C
C      do 1000 ki=1,3

```

```

do 1000 k2=1,3
  do 1000 k3=1,3
    do 1000 k4=1,3
      do 1000 k5=1,3
        do 1000 k6=1,3
          do 1000 k7=1,3
            do 1000 k8=1,3
              do 1000 k9=1,3
                do 1000 k10=1,3
do 1000 k11=1,3
  do 1000 k12=1,3
    do 1000 k13=1,3
      do 1000 k14=1,3
        do 1000 k15=1,3
          do 1000 k16=1,3
            do 1000 k17=1,3
              do 1000 k18=1,3
                do 1000 k19=1,3
                  do 1000 k20=1,3

c
c
      nbase(1)=k1
      nbase(2)=k2
      nbase(3)=k3
      nbase(4)=k4
      nbase(5)=k5
      nbase(6)=k6
      nbase(7)=k7
      nbase(8)=k8
      nbase(9)=k9
      nbase(10)=k10
      nbase(11)=k11
      nbase(12)=k12
      nbase(13)=k13
      nbase(14)=k14
      nbase(15)=k15
      nbase(16)=k16
      nbase(17)=k17
      nbase(18)=k18
      nbase(19)=k19
      nbase(20)=k20

c
c
do 1250 nn=1,jj
  n=0
  do 1200 j=1,nsup
    if(mset(nn,j).eq.1 .and. nbase(j).ne.1 .or.
1      mset(nn,j).eq.2 .and. nbase(j).ne.2 .or.
2      mset(nn,j).eq.3 .and. nbase(j).ne.3 .or.
3      mset(nn,j).eq.4 .and. nbase(j).ne.4) then
      n=n+1
    endif
1200    continue
  c
  c
  if(n.lt.ndiff) then
    goto 1000
  endif
1250 continue
  c
  c
  jj=jj+1
  write(*,130)(nbase(i),i=1,nsup),jj
  do 1100 i=1,nsup

```



```

                                mset(jj,i)=nbase(i)
1100                                continue
c
c
1000    continue
c
c
                                write(*,*)
130                                format(10x,20(1x,i1),5x,i5)
                                write(*,*)
                                write(*,120) jj
120                                format(1x,'Number of words=',i5)
c
c
                                end
c
c
                                *****
c
                                *****
c
```

APPENDIX Ic
Exemplary computer program for generating
minimally cross hybridizing sets
(double stranded tag/single stranded tag complement)

```

Program 3tagN
C
C
C      Program 3tagN generates minimally cross-hybridizing
C      sets of duplex subunits given i) N--subunit length,
C      and ii) an initial homopurine sequence.
C
C      character*1 sub1(20)
C      integer*2 mset(10000,20), nbase(20)
C
C
C      write(*,*) 'ENTER SUBUNIT LENGTH'
C      read(*,100) nsub
100    format(i2)
C
C
C      write(*,*) 'ENTER SUBUNIT SEQUENCE a & g only'
C      read(*,110) (sub1(k), k=1, nsub)
110    format(20a1)
C
C      ndiff=10
C
C      Let a=1 and g=2
C
C      do 800 kk=1, nsub
C      if(sub1(kk).eq.'a') then
C      mset(1, kk)=1
C      endif
C      if(sub1(kk).eq.'g') then
C      mset(1, kk)=2
C      endif
800    continue
C
C      jj=1
C
C      do 1000 k1=1, 3
C      do 1000 k2=1, 3
C      do 1000 k3=1, 3
C      do 1000 k4=1, 3
C      do 1000 k5=1, 3
C      do 1000 k6=1, 3
C      do 1000 k7=1, 3
C      do 1000 k8=1, 3
C      do 1000 k9=1, 3
C      do 1000 k10=1, 3
C      do 1000 k11=1, 3
C      do 1000 k12=1, 3
C      do 1000 k13=1, 3
C      do 1000 k14=1, 3
C      do 1000 k15=1, 3
C      do 1000 k16=1, 3
C      do 1000 k17=1, 3
C      do 1000 k18=1, 3

```

```

do 1000 k19=1,3
do 1000 k20=1,3
c
nbase(1)=k1
nbase(2)=k2
nbase(3)=k3
nbase(4)=k4
nbase(5)=k5
nbase(6)=k6
nbase(7)=k7
nbase(8)=k8
nbase(9)=k9
nbase(10)=k10
nbase(11)=k11
nbase(12)=k12
nbase(13)=k13
nbase(14)=k14
nbase(15)=k15
nbase(16)=k16
nbase(17)=k17
nbase(18)=k18
nbase(19)=k19
nbase(20)=k20
c
do 1250 nn=1,jj
c
n=0
do 1200 j=1,nsup
if(mset(nn,j).eq.1 .and. nbase(j).ne.1 .or.
1 mset(nn,j).eq.2 .and. nbase(j).ne.2 .or.
2 mset(nn,j).eq.3 .and. nbase(j).ne.3 .or.
3 mset(nn,j).eq.4 .and. nbase(j).ne.4) then
n=n+1
endif
1200 continue
c
if(n.lt.ndiff) then
goto 1000
endif
1250 continue
c
jj=jj+1
write(*,130) (nbase(i),i=1,nsup),jj
do 1100 i=1,nsup
mset(jj,i)=nbase(i)
1100 continue
c
1000 continue
c
write(*,*)
130 format(10x,20(1x,i1),5x,i5)
write(*,*)
write(*,120) jj
120 format(1x,'Number of words=',i5)
c
c
end

```

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i) APPLICANT: David W. Martin, Jr.

(ii) TITLE OF INVENTION: Measurement of Gene Expression profiles in Toxicity Determination

(iii) NUMBER OF SEQUENCES: 7

(iv) CORRESPONDENCE ADDRESS:

(A) ADDRESSEE: Stephen C. Macevicz, Lynx Therapeutics, Inc.
(B) STREET: 3832 Bay Center Place
(C) CITY: Hayward
(D) STATE: California
(E) COUNTRY: USA
(F) ZIP: 94545

(v) COMPUTER READABLE FORM:

(A) MEDIUM TYPE: 3.5 inch diskette
(B) COMPUTER: IBM compatible
(C) OPERATING SYSTEM: Windows 3.1
(D) SOFTWARE: Microsoft Word 5.1

(vi) CURRENT APPLICATION DATA:

(A) APPLICATION NUMBER:
(B) FILING DATE:
(C) CLASSIFICATION:

(vii) PRIOR APPLICATION DATA:

(A) APPLICATION NUMBER: PCT/US96/09513
(B) FILING DATE: 06-JUN-96

(vii) PRIOR APPLICATION DATA:

(A) APPLICATION NUMBER: PCT/US95/12791
(B) FILING DATE: 12-OCT-95

(viii) ATTORNEY/AGENT INFORMATION:

(A) NAME: Stephen C. Macevicz
(B) REGISTRATION NUMBER: 30,285
(C) REFERENCE/DOCKET NUMBER: 813wo

(ix) TELECOMMUNICATION INFORMATION:

(A) TELEPHONE: (510) 670-9365
(B) TELEFAX: (510) 670-9302

(2) INFORMATION FOR SEQ ID NO: 1:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 11 nucleotides
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 1:

CTAGTCGACC A

11

(2) INFORMATION FOR SEQ ID NO: 2:

(i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 11 nucleotides
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 2:

NRRGATCYNN N

11

(2) INFORMATION FOR SEQ ID NO: 3:

(i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 38 nucleotides
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 3:

GAGGATGCCT TTATGGATCC ACTCGAGATC CCAATCCA

38

(2) INFORMATION FOR SEQ ID NO: 4:

(i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 20 nucleotides
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: double
 (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 4:

AGTGGCTGGG CATCGGACCG

20

(2) INFORMATION FOR SEQ ID NO: 5:

(i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 20 nucleotides
 (B) TYPE: nucleic acid

(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 5:

GGGGCCCACT CAGCGTCGAT

20

(2) INFORMATION FOR SEQ ID NO: 6:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 nucleotides
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 6:

ATCGACGCTG ACTGGGCCCC

16

(2) INFORMATION FOR SEQ ID NO: 7:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 62 nucleotides
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 7:

AAAAGGAGGA GGCCTTGATA GAGAGGACCT GTTTAAACGG ATCCTCTTCC
TCTTCCTCTT CC

50

62

I claim:

1. A method of determining the toxicity of a compound, the method comprising the steps of:
 - 5 administering the compound to a test organism;
extracting a population of mRNA molecules from each of one or more tissues of the test organism;
forming a separate population of cDNA molecules from each population of mRNA molecules from the one or more tissues such that each cDNA molecule of a
10 separate population has an oligonucleotide tag attached, the oligonucleotide tags being selected from the same minimally cross-hybridizing set;
separately sampling each population of cDNA molecules such that substantially all different cDNA molecules within a separate population have different oligonucleotide tags attached;
15 sorting the cDNA molecules of each separate population by specifically hybridizing the oligonucleotide tags with their respective complements, the respective complements being attached as uniform populations of substantially identical complements in spatially discrete regions on one or more solid phase supports;
determining the nucleotide sequence of a portion of each of the sorted cDNA
20 molecules of each separate population to form a frequency distribution of expressed genes for each of the one or more tissues; and
correlating the frequency distribution of expressed genes in each of the one or more tissues with the toxicity of the compound.
- 25 2. The method of claim 1 wherein said oligonucleotide tag and said complement of said oligonucleotide tag are single stranded.
3. The method of claim 2 wherein said oligonucleotide tag consists of a plurality of subunits, each subunit consisting of an oligonucleotide of 3 to 9 nucleotides in
30 length and each subunit being selected from the same minimally cross-hybridizing set.
4. The method of claim 3 wherein said one or more solid phase supports are microparticles and wherein said step of sorting said cDNA molecules onto the microparticles produces a subpopulation of loaded microparticles and a subpopulation
35 of unloaded microparticles.
5. The method of claim 4 further including a step of separating said loaded microparticles from said unloaded microparticles.

6. The method of claim 5 further including a step of repeating said steps of sampling, sorting, and separating until a number of said loaded microparticles is accumulated is at least 10,000.
- 5
7. The method of claim 6 wherein said number of loaded microparticles is at least 100,000.
8. The method of claim 7 wherein said number of loaded microparticles is at least 500,000.
- 10
9. The method of claim 5 further including a step of repeating said steps of sampling, sorting, and separating until a number of said loaded microparticles is accumulated is sufficient to estimate the relative abundance of a cDNA molecule present in said population at a frequency within the range of from 0.1% to 5% with a 95% confidence limit no larger than 0.1% of said population.
- 15
10. The method of claim 4 wherein said test organism is a mammalian tissue culture.
- 20
11. The method of claim 10 wherein said mammalian tissue culture comprises hepatocytes.
12. The method of claim 4 wherein said test organism is an animal selected from the group consisting of rats, mice, hamsters, guinea pigs, rabbits, cats, dogs, pigs, and monkeys.
- 25
13. The method of claim 12 wherein said one or more tissues are selected from the group consisting of liver, kidney, brain, cardiovascular, thyroid, spleen, adrenal, large intestine, small intestine, pancreas urinary bladder, stomach, ovary, testes, and mesenteric lymph nodes.
- 30
14. A method of identifying genes which are differentially expressed in a selected tissue of a test animal after treatment with a compound, the method comprising the steps of:
- 35
- administering the compound to a test animal;

extracting a population of mRNA molecules from the selected tissue of the test animal;

forming a population of cDNA molecules from the population of mRNA molecules such that each cDNA molecule has an oligonucleotide tag attached, the oligonucleotide tags being selected from the same minimally cross-hybridizing set;

sampling the population of cDNA molecules such that substantially all different cDNA molecules have different oligonucleotide tags attached;

sorting the cDNA molecules by specifically hybridizing the oligonucleotide tags with their respective complements, the respective complements being attached as uniform populations of substantially identical complements in spatially discrete regions on one or more solid phase supports;

determining the nucleotide sequence of a portion of each of the sorted cDNA molecules to form a frequency distribution of expressed genes; and

identifying genes expressed in response to administering the compound by comparing the frequency distribution of expressed genes of the selected tissue of the test animal with a frequency distribution of expressed genes of the selected tissue of a control animal.

15. The method of claim 14 wherein said oligonucleotide tag and said complement of said oligonucleotide tag are single stranded.

16. The method of claim 15 wherein said oligonucleotide tag consists of a plurality of subunits, each subunit consisting of an oligonucleotide of 3 to 9 nucleotides in length and each subunit being selected from the same minimally cross-hybridizing set.

17. The method of claim 16 wherein said one or more solid phase supports are microparticles and wherein said step of sorting said cDNA molecules onto the microparticles produces a subpopulation of loaded microparticles and a subpopulation of unloaded microparticles.

18. The method of claim 17 further including a step of separating said loaded microparticles from said unloaded microparticles.

19. The method of claim 18 further including a step of repeating said steps of sampling, sorting, and separating until a number of said loaded microparticles is accumulated is at least 10,000.

20. The method of claim 19 wherein said number of loaded microparticles is at least 100,000.

21. The method of claim 20 wherein said number of loaded microparticles is at least 500,000.

22. The method of claim 18 further including a step of repeating said steps of sampling, sorting, and separating until a number of said loaded microparticles is accumulated is sufficient to estimate the relative abundance of a cDNA molecule present in said population at a frequency within the range of from 0.1% to 5% with a 95% confidence limit no larger than 0.1% of said population.

23. The method of claim 17 wherein said test animal is selected from the group consisting of rats, mice, hamsters, guinea pigs, rabbits, cats, dogs, pigs, and monkeys.

24. The method of claim 23 wherein said selected tissue is selected from the group consisting of liver, kidney, brain, cardiovascular, thyroid, spleen, adrenal, large intestine, small intestine, pancreas urinary bladder, stomach, ovary, testes, and mesenteric lymph nodes.

25. A use of the technique of massively parallel signature sequencing to determine the toxicity of a compound in a test organism, the use comprising the steps of:

administering the compound to a test organism;

extracting a population of mRNA molecules from each of one or more tissues of the test organism and forming a population of cDNA molecules for each of the one or more tissues;

determining the nucleotide sequence of a portion of each of the cDNA molecules of each separate population using massively parallel signature sequencing to form a frequency distribution of expressed genes for each of the one or more tissues; and

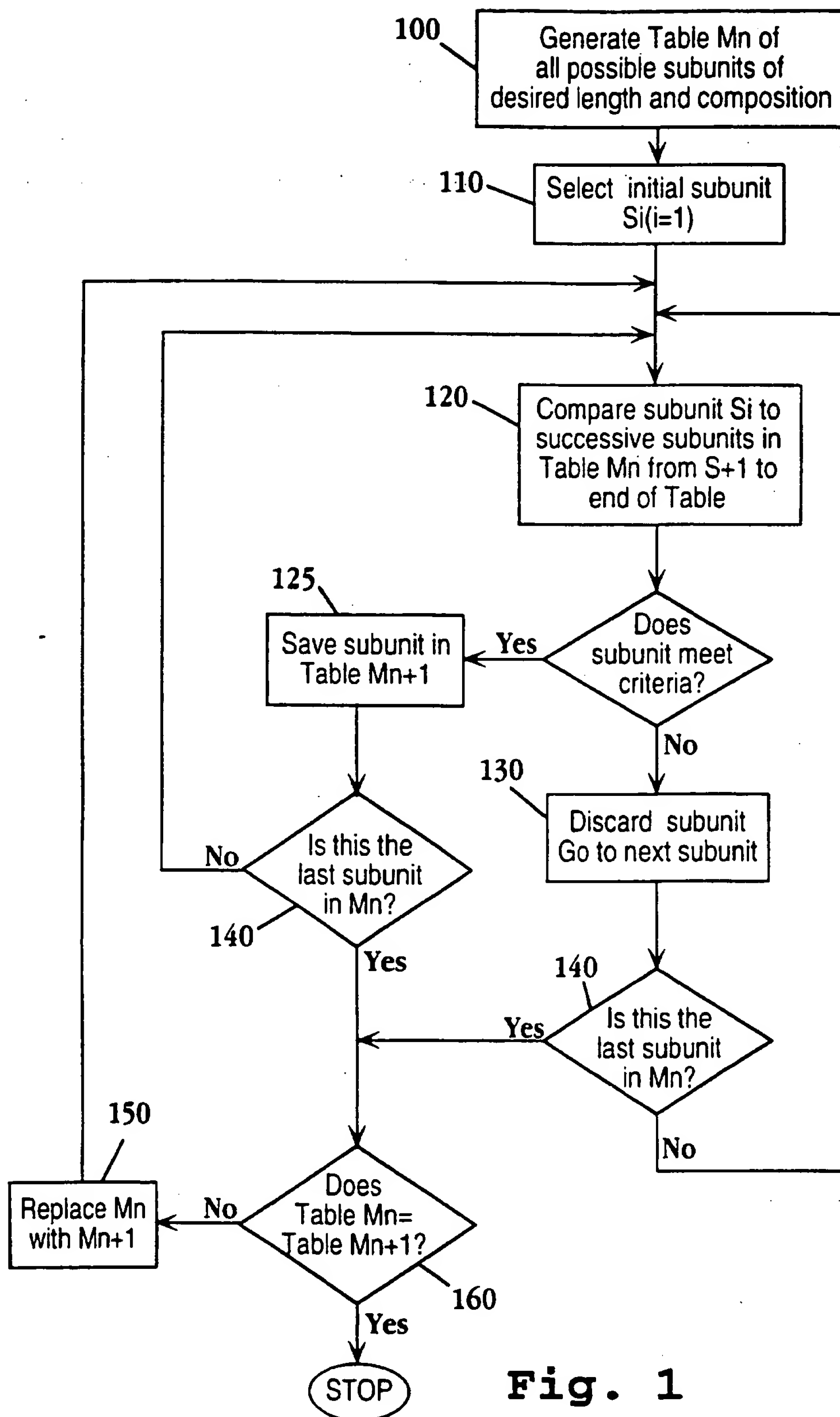
correlating the frequency distribution of expressed genes in each of the one or more tissues with the toxicity of the compound.

26. The use of claim 25 wherein said test organism is a mammalian tissue culture.

27. The use of claim 26 wherein said mammalian tissue culture comprises hepatocytes.

28. The use of claim 25 wherein said test organism is an animal selected from the group consisting of rats, mice, hamsters, guinea pigs, rabbits, cats, dogs, pigs, and monkeys.
- 5 29. The use of claim 28 wherein said one or more tissues are selected from the group consisting of liver, kidney, brain, cardiovascular, thyroid, spleen, adrenal, large intestine, small intestine, pancreas urinary bladder, stomach, ovary, testes, and mesenteric lymph nodes.
- 10 30. A use of the technique of massively parallel signature sequencing to identify genes which are differentially expressed in a test organism after treatment with a compound and which are correlated with toxicity of the compound, the use comprising the steps of:
- 15 administering the compound to the test organism;
extracting a population of mRNA molecules from a selected tissue of the test organism and forming a population of cDNA molecules;
determining the nucleotide sequence of a portion of each of the cDNA molecules using massively parallel signature sequencing to form a frequency distribution of expressed genes;
- 20 identifying genes expressed in response to administering the compound by comparing the frequency distribution of expressed genes of the selected tissue of the test organism with a frequency distribution of expressed genes of the selected tissue of a control organism; and
- 25 determining whether the genes expressed in response to administering the compound are correlated with toxicity of the compound in the test organism.

1/2

**Fig. 1**

2/2

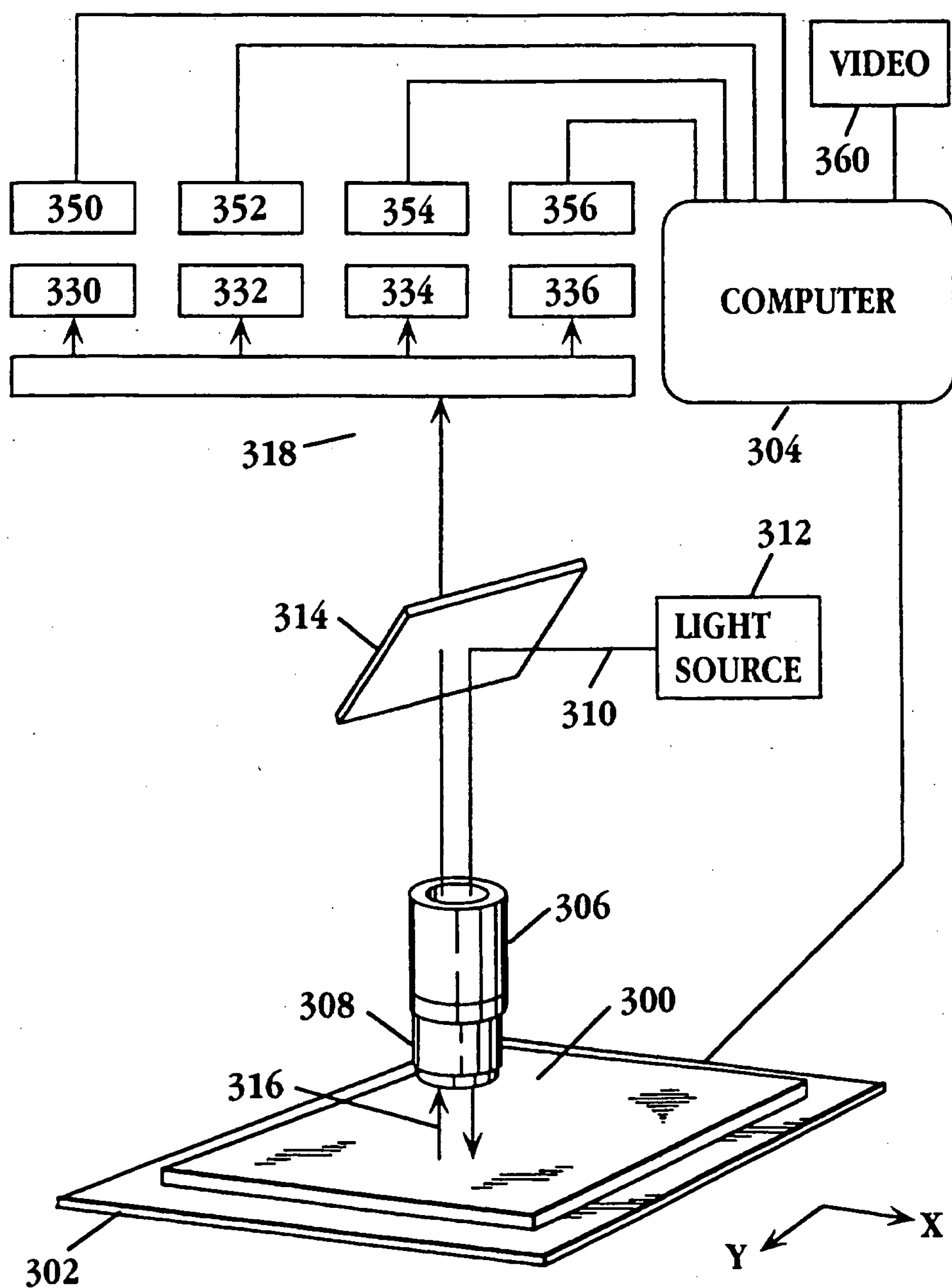


Fig. 2

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US96/16342**A. CLASSIFICATION OF SUBJECT MATTER**

IPC(6) : C12Q 1/68; C07H 21/04

US CL : 435/6; 536/24.3

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6; 536/24.3

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS, MEDLINE, BIOSIS, CAPLUS, SCISEARCH

search terms: Martin, David W., toxic?, differential?, express?, cDNA, mRNA, RNA, gene#, hybrid?

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CHETVERIN et al. Oligonucleotide arrays: New concepts and possibilities. Bio/Technology. 12 November 1994, Vol. 12, pages 1093-1099, especially pages 1095-1096.	1-30
A	BRENNER et al. Encoded combinatorial chemistry. Proceedings of the National Academy of Sciences USA. June 1992, Vol. 89, pages 5381-5383.	1-30
A	MATSUBARA et al. cDNA analyses in the human genome project. Gene. 15 December 1993, Vol. 135, No. 1-2, pages 265-274.	1-30



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:	*T	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
A document defining the general state of the art which is not considered to be of particular relevance	*X*	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
E earlier document published on or after the international filing date	*Y*	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
L document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	*G*	document member of the same patent family
O document referring to an oral disclosure, use, exhibition or other means		
P document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search

27 JANUARY 1997

Date of mailing of the international search report

19 FEB 1997

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer

SCOTT D. PRIEBE

Telephone No. (703) 308-0196

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US96/16342

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 95/21944 A1 (SMITHKLINE BEECHAM CORPORATION) 17 August 1995, page 4, lines 1-4, page 5, lines 31-37, page 17, lines 15-27, page 18, lines 30-35, page 20, line 23 to page 21, line 4.	1-30

FOCUS - 17 of 19 DOCUMENTS

Copyright 1997 PR Newswire Association, Inc.
PR Newswire

August 11, 1997, Monday

SECTION: Financial News

DISTRIBUTION: TO BUSINESS AND MEDICAL EDITORS

LENGTH: 478 words

HEADLINE: Eli Lilly & Co. and Acacia Biosciences Enter Into Research Collaboration;
First Corporate Agreement for Acacia's Genome Reporter Matrix(TM)

DATELINE: RICHMOND, Calif., Aug. 11

BODY:

Acacia Biosciences and Eli Lilly and Company (Lilly) announced today the signing of a joint research collaboration to utilize Acacia's Genome Reporter Matrix(TM) (GRM) to aid in the selection and optimization of lead compounds. Under the collaboration, Acacia will provide chemical and biological profiles on a class of Lilly's compounds for an undisclosed fee.

Acacia's GRM is an assay-based computer modeling system that uses yeast as a miniature ecosystem. The GRM can profile the extent, nature and quantity of any changes in gene expression. Because of the similarities between the yeast and human genome, the system serves as an excellent surrogate for the human body, mimicking the effects induced by a biologically active molecule.

"Using yeast as a model organism for lead optimization makes a lot of sense given the high degree of homology with human metabolic pathways," said William Current of Lilly Research Laboratories. "Acacia's innovative GRM has the potential to provide enormous insight into the therapeutic impact of our compounds and make the drug discovery process more rational. It should substantially accelerate the development process."

"This first agreement with a major pharmaceutical company is an important milestone in the development of Acacia," said Bruce Cohen, President and CEO of Acacia. "The deal is in line with our strategy of establishing alliances that will allow our collaborators to use genomic profiles to identify and optimize compounds within their existing portfolios. In the long run, this technology can be used to characterize large scale combinatorial libraries, predict side effects prior to clinical trials and resurrect drugs that have failed during clinical trials."

The GRM incorporates two critical elements: chemical response profiles and genetic response profiles. The chemical response profiles measure the change in gene expression caused by potential therapeutics and then rank genes with altered expressions by degree of response. The genetic response profiles measure changes in gene expression caused by mutations in the genes encoding potential targets of pharmaceuticals; these genetic response profiles represent gold standards in drug discovery by defining the response profile expected for drugs with perfect selectivity and specificity. By comparing the two profiles, one can analyze a potential drug candidate's ability to mimic the action of a 'perfect' drug.

Acacia Biosciences is a functional genomics company developing proprietary technologies to enhance the speed and efficacy of drug discovery and development. Acacia's Genome Reporter Matrix capitalizes on the latest advances in genomics and combinatorial chemistry to generate comprehensive profiles of drug candidates' in vivo activity.

SOURCE Acacia Biosciences

CONTACT: Bruce Cohen, President and CEO of Acacia Biosciences, 510-669-2330 ext. 103 or Media: Linda Seaton of Feinstein

LOAD-DATE: August 12, 1997

The Bioreactor Market:
Steady Growth Expected

The worldwide market for all bioreactors was valued at \$275 million for 1997, and is expected to be worth \$380 million by 2002.

1997 2002
\$275 million \$380 million

V.17
C.01
TI: GENETIC ENGINEERING NEWS

W1 GE281M
NO.16
1897
SEQ: G04575000

08/25/97

GENETIC ENGINEERING NEWS

GEN

BIOTECHNOLOGY • BIOPROCESS • BIORESEARCH • TECHNOLOGY TRANSFER

Contents	
European Biotech Standards Moving	4
Bioprocess Simulation Packages	13
Trends in Biotechnology Development	14
QC/QA for Small Biotech Firms	16
Advances in Electroporation	19
New Products	21
New GEN Column—Drug Discovery	27
Corporate Profiles: Pangen Systems	28
Canada Watch	29
European Roundup	30
Wall Street Outlook	31
Collaboration Agreements	32
Trade Industry	33
Canada Trade Update	37
Hot Companies	40
People	41
Calendar	41
Marketplace	42

Pharmagene Raises More Capital for Research on Human Tissues

By Sophia Fox

Pharmagene, the Royston, U.K.-based biopharmaceutical company specializing in the use of human biomaterials for drug discovery research, has raised a further £5 million from a group of investors led by 3i and Abacus Nominees. The funding will enable the company to expand both its human biomaterials collection and its capabilities across a range of proprietary platform technologies.

Gordon Baxter, Ph.D., Pharmagene's cofounder and chief operating officer, claimed, "by the end of this year Pharmagene will have access to the largest collection of human RNAs and proteins anywhere in the world, and a range of innovative, yet robust technologies."

SEE PHARMAGENE, P. 9

Perkin-Elmer Acquires PerSeptive to Expand Its Capabilities in Gene-Based Drug Discovery

By John Sterling

Perkin-Elmer's (PE; Norwalk, CT) decision last month to acquire PerSeptive Biosystems (Framingham, MA) via a \$360 million stock swap was designed to strengthen PE in terms of broad capabilities in gene-based drug discovery. The company's main goal is to develop new products to improve the integration of genetic and protein research.

"This merger will enhance our position as an effective provider of innovative, integrated platforms enabling our customers to be more efficient and cost-effective in bringing new pharmaceuticals to market," says Tony L. White, PE's chairman, president and CEO. "The combination of our two companies should bolster our presence in the life sciences, [and it is our] belief that we must take bold action now to lead the emerging era of molecular medicine with leading positions in both genetic and protein analysis."

A driving force behind the merger is the vast amount of genetic



Perkin-Elmer acquired PerSeptive Biosystems for \$360 million to obtain new technologies in mass spectrometry, bioseparations and purification for product development projects, spanning the range from genomics to proteomics.

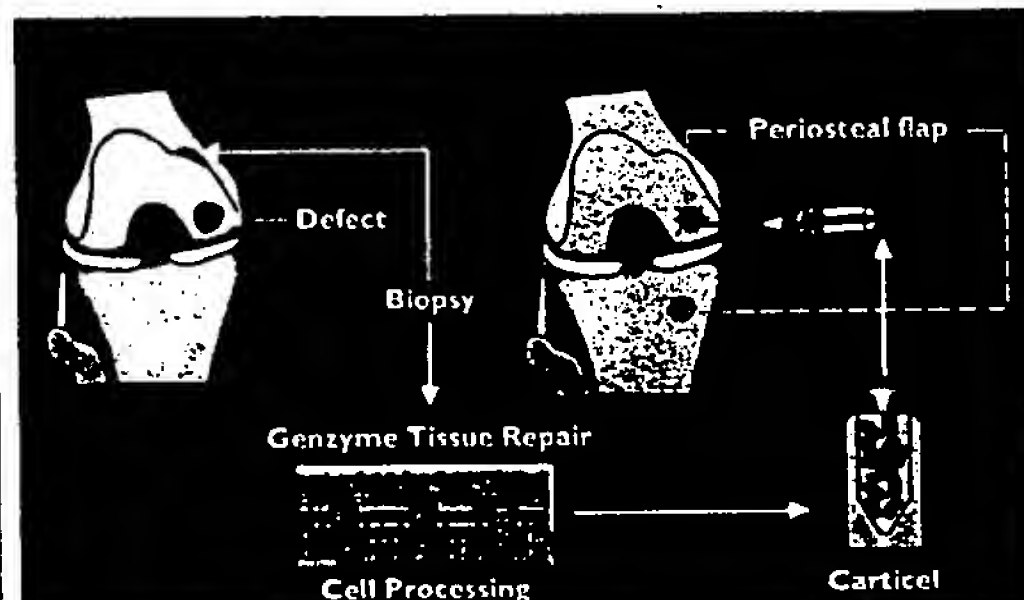
information about human disease that is being accumulated by researchers and biotech companies working in the area of genomics. It is becoming increasingly obvious that these data need to be complemented with technologies for

studying proteins and protein networks—a field known as proteomics (see GEN, September 1, 1997, p. 1).

PE officials, who claim that MALDI-TOF (Matrix Assisted Laser Desorption/Ionization) is a key technology for

SEE ACQUISITION, P. 10

FDA OKs Genzyme's Carticel Product for Damage to Knees



Carticel, which was approved for the repair of clinically significant, symptomatic cartilaginous defects of the femoral condyle (medial, lateral or trochlear) caused by acute or repetitive trauma, employs a proprietary process to grow autologous cartilage cells for implantation.

By Naomi Pfeiffer

The FDA has approved a knee-cartilage replacement product made by Genzyme Tissue Repair (Cambridge, MA), a tracking-stock division of Genzyme Corp., for people with trauma-damaged knees.

Carticel (autologous cultured chondrocytes) is the first product to be licensed under the FDA's process to grow autologous cartilage cells for implantation.

SEE GENZYME, P. 6

Strategies for Target Validation Streamline Evaluation of Leads

By Vicki Glaser

Acacia Biosciences (Richmond, CA) last month announced its first agreement with a major pharmaceutical company, signing a deal with Eli Lilly (Indianapolis, IN) to use Acacia's Genome Reporter Matrix (GRM) to select and optimize some of Lilly's lead compounds. Acacia's yeast-based system for profiling drug activity is useful for evaluating the therapeutic potential of lead compounds, and it also has a role in the identification and validation of new drug targets.

"We're using the ecosystem of a cell to allow us to deduce the mechanism of action and target for any chemical," explains Bruce Cohen, president and CEO. "We screen for every target in a cell simultaneously...using transcription as a readout

for how a cell is adapting to any perturbation," he says.

The GRM technology consists of two main databases: one is the genetic response profile, showing the effects of mutations in each individual yeast gene and compensatory gene regulatory mechanisms; the other is the chemical response profile, which documents changes in gene expression in response to chemical compounds. Computational analysis and pattern matching between the genetic and chemical profiles yields information on the specificity, potency and side-effects risk of a drug lead.

Targeting Targets

No longer is mapping and sequencing a gene—or the human genome—an end unto itself, but

SEE TARGET, P. 15

Sticky Ends

Avigen received two grants from the NIH & University of California for research on gene therapy for treatment of cancer & HIV infections...MRL Pharmaceutical Services, of Reston, VA, launched the TSN Bug Finder, which is able to locate & retrieve client-specified microorganisms in real-time...Gensia Sidor, Inc. will move its corporate staff from San Diego to Irvine, CA, by end of year...

FDA accepted NDA from Sepracor for levalbuterol HCl inhalation solution...An \$11.7M mezzanine financing has been closed by Activated Cell Therapy, which changed its name to Dendreon Corporation...Astra AB will build major research facility in Waltham, MA, and is also relocating Astra Arcus research facility from Rochester to Boston area...Prolifix Ltd. team used a small peptide to inhibit the E2F protein complex and induced

apoptosis in mammalian tumor cells...Vertex Pharmaceuticals, Inc. and Alpha Therapeutic Corp. ended an agreement to develop VX-366 for treatment of inherited hemoglobin disorders...NaviCyte received Phase I SBIR grant for up to \$100,000 from NIH for development of prototype of its NaviFlow technology for high-throughput screening...Covance Inc. will invest \$21 million in expansion and renovation of its facility in Indianapolis, IN.



Target

from page 1

merely a means to an end. The critical next step is to validate the gene and its protein product as a potential drug target. The Human Genome Project continues to produce a treasure chest of expressed sequence tags (ESTs) and a tantalizing array of complete gene sequences.

Companies are applying a variety of functional genomic strategies to link genes to specific diseases and to multigenic phenotypes. Yet the ultimate challenge for pharmaceutical companies is to sift through all the sequence and differential gene expression data to identify the best targets for drug discovery.

Spinning off technology developed at the University of North Carolina (Chapel Hill), Cytogen Corp. (Princeton, NJ) formed its wholly owned subsidiary AxCell Biosciences earlier this year. The young company is building a protein interaction database, cataloging all the interactions the modular domains of proteins can engage in with a

range of ligands, in order to gain insight into protein function and to select the most critical interaction to target for drug development.

AxCell's cloning-of-ligand-targets (COLT) technology employs "recognition units" from the company's genetic diversity library (GDL) to map functional protein interactions and quantitate their affinity. The company's inter-functional proteomic database (IFP-dbase) elucidates protein interaction networks and structure-activity relationships based on ligand affinity with protein modular domains.

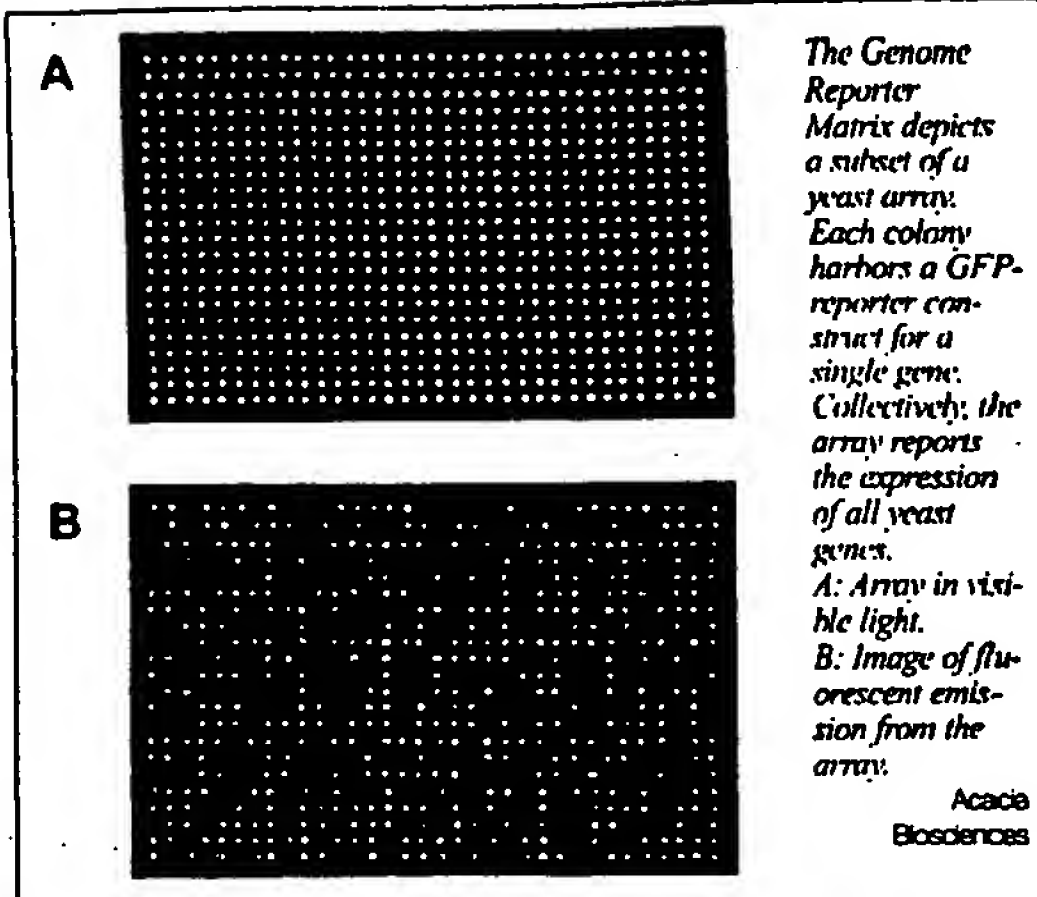
Defining Disease Pathways

Signal Pharmaceuticals, Inc.'s (San Diego, CA) integrated drug target and discovery effort is based on mapping gene-regulating pathways in cells and identifying small molecules that regulate the activation of those genes. In collaboration with academic researchers, the company has identified a large number of regulatory proteins in several mitogen-activated protein (MAP) kinase pathways (including the JNK, FRK and p38

signaling pathways), which Signal is evaluating for the treatment of autoimmune, inflammatory, cardiovascular and neurologic diseases, and cancer. Other target identification

programs focus on the NF- κ B pathway, estrogen-related genes and central/peripheral nervous system genes.

Regulating cytokine production in immune and inflammatory disorders,



The Genome Reporter Matrix depicts a subset of a yeast array. Each colony harbors a GFP-reporter construct for a single gene. Collectively, the array reports the expression of all yeast genes.
A: Array in visible light.
B: Image of fluorescent emission from the array.

Acacia Biosciences

and modifying bone metabolism to treat osteoporosis are the focus of Signal's collaboration with Tanabe Selyaku (Osaka, Japan). Signal has partnered with Organon/Akzo Nohel (Netherlands) to identify estrogen-responsive genes as targets for treating neurodegenerative and psychiatric diseases, atherosclerosis and ischemia, and with Roche Bioscience (Palo Alto, CA) to develop human peripheral nerve cell lines for the discovery of treatments for pain and incontinence.

Exelixis (S. San Francisco, CA) strategy for target selection is to define disease pathways and identify regulatory molecules that activate or inhibit those biochemical/genetic pathways. Based on the finding that these pathways are conserved across species, the company is studying the model genetic systems of *Drosophila* and *Caenorhabditis elegans*. Using its PathFinder technology, Exelixis systematically introduces mutations into the genomes of these model organisms, looking for mutations that enhance or suppress the target disease-related gene. These novel genes then become the basis of drug screening assays.

Cadus Pharmaceutical Corp. (Tarrytown, NY) is identifying surrogate ligands to newly discovered orphan G-protein coupled transmembrane receptors of unknown function to determine the suitability of the receptors as drug targets. Inserting the novel receptor in a yeast system yields a ligand that activates the receptor. Access to a surrogate ligand allows the company to screen for receptor antagonists in the yeast system.

"The antagonist plus the surrogate ligand gives you two probes—an on probe and an off probe—which allows you to look at function," explains David Webb, Ph.D., vp of research and chief scientific officer. A surrogate ligand also provides information on which G-protein interacts with the orphan receptor and its associated signaling pathways, further clarifying the role of the receptor as a potential drug target. Cadus' collaboration with SmithKline (Philadelphia) capitalizes on Cadus' ability to determine orphan receptor function, applying the technology to SmithKline's proprietary, newly discovered G-protein receptors.

Cadus' recombinant yeast system can also be used to screen cell and tissue extracts for natural ligands, and the company is accelerating its internal drug-discovery efforts in the areas of cancer, inflammation and allergy. A recent equity investment in Axiom Biotechnologies (San Diego, CA) gave Cadus a license to Axiom's high-throughput pharmacologic screening system for lead optimization and discovery.

As its name implies, gene/Networks (Alameda, CA) focuses on identifying gene networks that contribute to multigenic phenotypes and complex disease processes. The integration of mouse and human genetic studies forms the basis of the technology. The Genome Tagged Mice database in development will serve as a library of natural mouse genetic and phenotypic variation. Disease-related genes identified in mice are then evaluated in human family- and population-based studies to confirm their clinical relevance and linkages to pathophysiologic traits.

Blocking Gene Expression

Inactivating a gene known to be expressed in association with a particular disease is one approach to identifying appropriate therapeutic targets. The target validation and discovery program at Ribozyme Pharmaceuticals, Inc. (Boulder, CO) applies the company's ribozyme technology to achieve selective inhibition of gene expression in cell culture and in animals.

Correlation of the gene expression inhibition with phenotype can

SEE TARGET, P. 38

A strong chemical combination to help you grow. And flourish.

Three hundred million dollars and ten years of hard work. That's what it costs to bring your biotechnology-derived therapeutic to the marketplace.

Which means, no room for error.

Which means, in turn, you'd be wise to tap into the combined capabilities of Mallinckrodt and J.T.Baker: dual sources, trusted names for your chemical raw materials.

Two separate GMP-produced brands offering the control of a single quality system and the convenience of a single audit process.

We offer comprehensive product lines including USP salts, bioreagents, high purity solvents and chromatography products in Beaker to Bulk™ packaging for easy scale-up.

Call 1-800-582-2537, or access our website at <http://www.mallinckrodt.com>. For dual chemical sources dedicated to helping you grow. Flourish. Succeed!

MALLINCKRODT



Target

from page 15

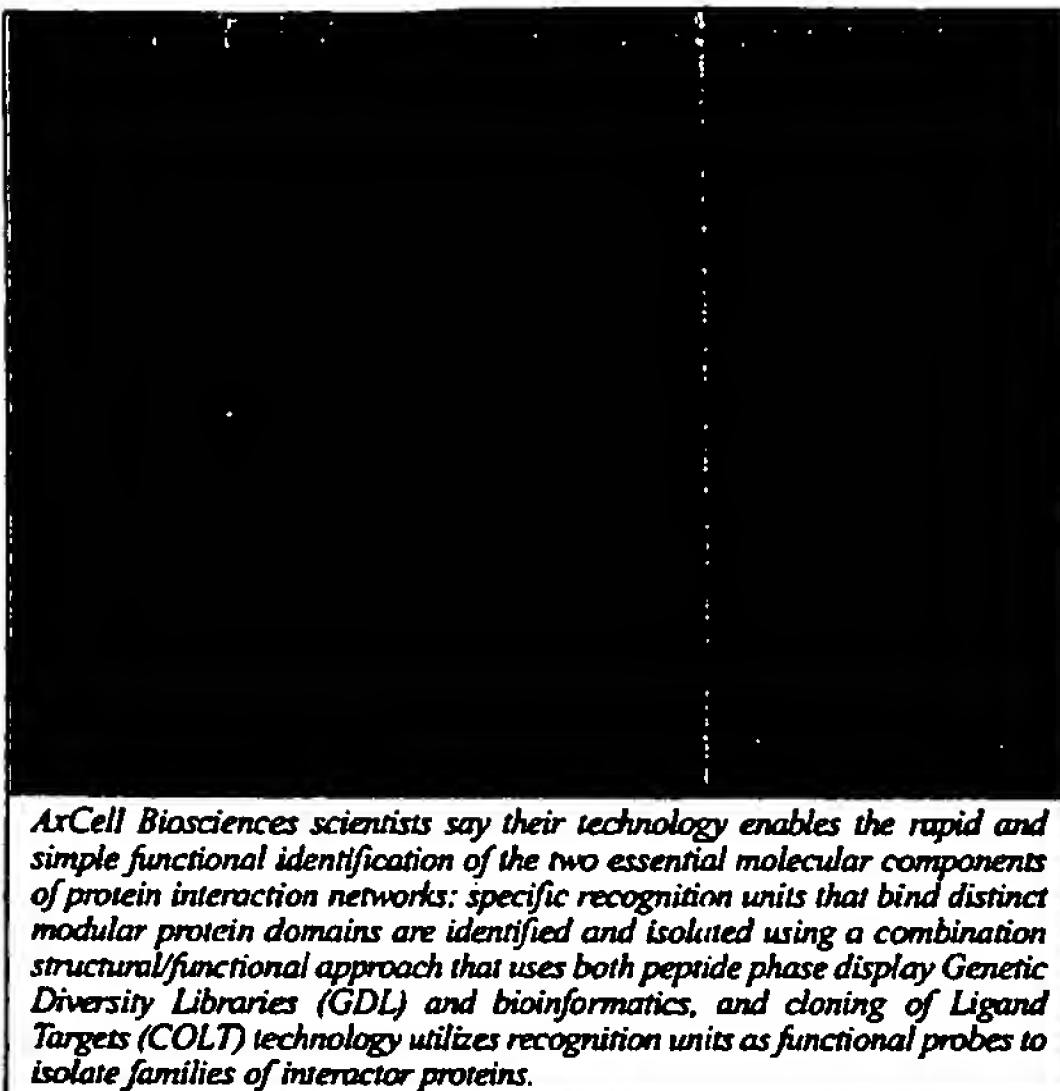
suggest the relative importance of the gene in disease pathology. The company's nuclease-resistant ribozymes form the basis of a collaboration with Schering AG (Germany) for drug target validation and the development of ribozyme-based therapeutic agents, and with Chiron Corp. (Emeryville, CA) for target validation.

With several antisense compounds now progressing through clinical trials, the concept of using oligonucleotides to inhibit gene activity is not new. But rather than focusing on therapeutics development, Sequitur, Inc. (Natick, MA) is creating antisense compounds for the purpose of determining gene function and validating drug targets. Clients typically provide the one-year-old company with the sequence (or EST) of a potential gene target and, in return, Sequitur custom designs a series of three to six antisense compounds that yield a three-to-ten-fold inhibition of the target gene in cell culture. The company also provides oligofectins, a series of cationic lipids, to deliver the oligonucleotides to a variety of cultured cells.

"Differential expression information is just for correlation, it doesn't tell function or confirm what would be a good target," says Tod Woolf, Ph.D., director of technology development at Sequitur. Whereas, antisense compounds will inhibit a target, Sequitur offers both phosphorothioate DNA antisense compounds, and its proprietary Next Generation chimeric oligonucleotides, which have a higher hybridization affinity, greater specificity and reduced toxicity, according to the company.

Mining Pathogen Genomes

Companies such as Human Genome Sciences (HGS; Rockville, MD), Incyte (Palo Alto, CA),



AxCell Biosciences scientists say their technology enables the rapid and simple functional identification of the two essential molecular components of protein interaction networks: specific recognition units that bind distinct modular protein domains are identified and isolated using a combination structural/functional approach that uses both peptide phase display Genetic Diversity Libraries (GDL) and bioinformatics, and cloning of Ligand Targets (COLT) technology utilizes recognition units as functional probes to isolate families of interactor proteins.

Millennium Pharmaceuticals Inc. (Cambridge, MA) and Genome Therapeutics (Waltham, MA) are relying on high-speed DNA sequencing, positional cloning and other strategies to identify specific microbial genomic sites that would be good targets for infectious disease therapeutics.

HGS recently completed sequencing of the bacterial pathogen *Streptococcus pneumoniae*, which is the focus of an agreement with Hoffmann-La Roche (Basel, Switzerland). Roche will use the sequence data to develop new anti-infectives against *S. pneumoniae*. HGS and Roche have expanded their collaboration to include a nonexclusive license to access sequence information for the intestinal bacterium *Enterococcus faecalis*.

Incyte Pharmaceuticals has completed one-fold coverage of the *Candida albicans* genome, identify-

ing 60% of the genes of this fungal pathogen. This genome will become part of the company's PathoSeq microbial database. Incyte recently introduced the ZooSeq animal gene sequence and expression database. The database will provide genomic information across various species commonly used in preclinical drug testing, which may help to better define potential drug targets.

Millennium Pharmaceuticals continues to report success in identifying novel drug targets, having recently discovered a novel chemokine called neurotactin and a new class of MAD-related proteins that inhibit transforming growth factor beta (TGF- β) signaling. The company also received U.S. patent coverage for the tub genes, believed to play a role in obesity, and for the gene that encodes the protein melastatin, which appears to suppress metastasis in malignant melanoma.

Pangea

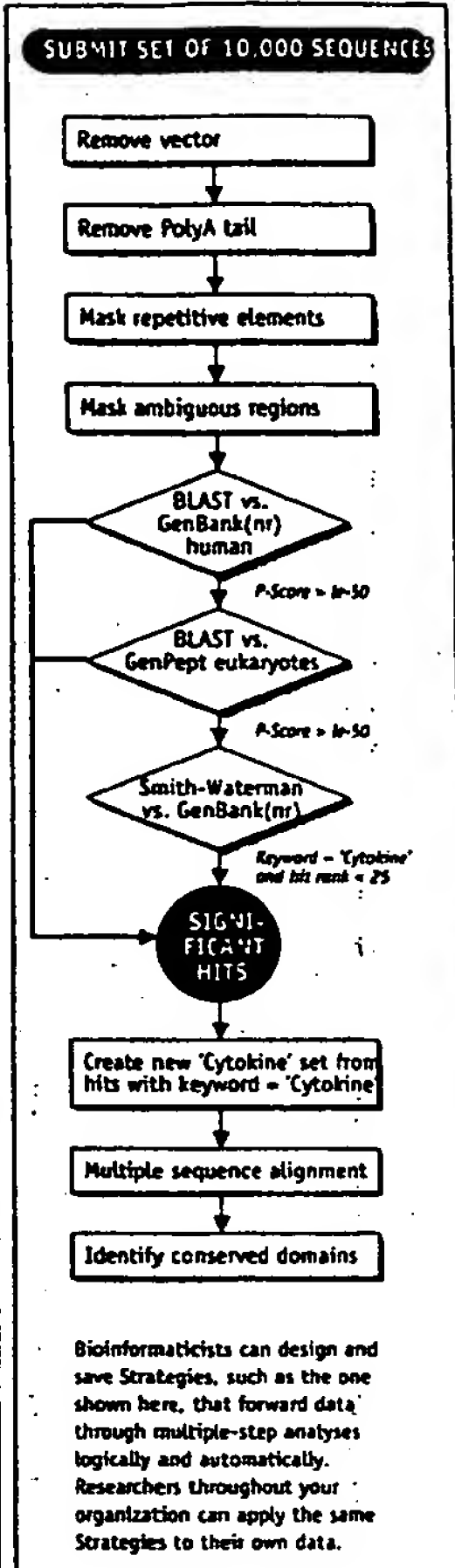
from page 28

Smith, now a computer programmer, is an expert in systems integration, Internet technologies and the application of industrial engineering principles to the drug discovery process. Before co-founding Pangea, he was the manager of software development at Attorney's Briefcase, a legal research software company.

By being "in the trenches" with customers and collaborators, Bellenson and Smith sensed the frustration of pharmaceutical researchers whose incompatible tools have impeded their progress. According to Bellenson, "Most of them are geared toward analyzing one molecule at a time. It's like emptying the ocean with an eye dropper—an incompatible eye dropper at that. A pharmaceutical company may have 30 different drug discovery teams with various approaches. The problem is to manage the process of experimenting with a lot of different approaches, to automate while maintaining flexibility."

GeneWorld 2.1 enables "integration of the entire target discovery and validation process," Bellenson says. The commercial software package coordinates the entire process of sequence-data analysis and can be integrated with other programs and databases, according to Smith, who adds that it handles thousands of sequence results, organizes and automates annotation and seamlessly interacts with growing genome databases. Simple forms and menus enable users to turn raw sequence data into crucial knowledge for drug discovery by applying algorithms to sequences, creating custom analysis strategies and producing useful reports, without the need for writing computer code. GeneWorld 2.1 runs on a variety of platforms and operating systems.

Pairing industrial relational database-management systems with a web-browser interface, Pangea's Operating System of Drug Discovery "is an open-computing framework that allows client/server and Java-enabled web-based technologies to collect, organize and analyze drug discovery information for pharmaceutical companies to simplify and accelerate drug discovery. The technology unites automated genomics database analysis for drug target site selection, chemical information database analysis and large-scale combinatorial chemistry project management and high-throughput screening project management for drug lead efficacy analysis. Pangea officials maintain that these integrated elements provide a unified environment for chemists, biologists and others involved in the drug discovery process to work together with



commercial and public domain software.

Pangea's Operating System of Drug Discovery can accommodate Sybase, Oracle or Informix relational database-management systems and any version of UNIX. It absorbs new data formats, databases, algorithms and analysis paradigms into the automated workflow without software modifications. Netscape Navigator provides a friendly user interface from PC, Macintosh, and UNIX workstations.

In the near term, Pangea plans to complete its bioinformatics core with two more programs. Gene Foundry, a sample tracking and workflow sequence package for DNA sequence and fragment information, will also offer interaction with robots, reagent tracking and troubleshooting. Gene Thesaurus, the other package is a "warehouse of bioinformatics data," says Bellenson.

Europe

from page 30

GTAC Chairman, Professor Norman C. Nevin, said 1996 saw "four important developments": an increase in enquiries and submissions made to GTAC; an increase in the complexity of submitted protocols; a continuing shift from gene therapy for single-gene disorders toward strategies aimed at tumour destruction in cancer; and a growth in international sponsorship of U.K. gene therapy trials.

Since 1993, GTAC and its predecessor, the Clothier Committee, have approved 18 U.K. gene therapy clinical trials (13 of which have been carried out), which are listed in the report. The disease areas targeted by these trials include severe combined immunodeficiency (1 trial), cystic fibrosis (6), metastatic melanoma (2), lymphoma (2), neuroblastoma (1), breast cancer (1), Hurler's syndrome (1), cervical cancer (1), glioblastoma

breast cancer, breast cancer with liver metastases, glioblastoma, malignant ascites due to gastrointestinal cancer and ovarian cancer.

Copies of the GTAC third annual report are available from the GTAC Secretariat, Wellington House, 133-155 Waterloo Road, London SE1 8UG, U.K.

Coated Lenses Prevent PCO

Scientists in the U.K. say it may be possible to prevent posterior capsule opacification (PCO), a common complication following cataract surgery, by using the implanted polymethylmethacrylate (PMMA) intraocular lens as a drug delivery system. PCO occurs in 30-50% of cataract surgery patients as a result of stimulated cell growth within the remaining capsular bag. The condition causes a decline in visual acuity and requires expensive laser treatment, thus negating the routine use of cataract surgery in underdeveloped countries, explains G. Duncan, at the



NEW HIGH SPECIFIC ACTIVITY MICROBIAL ALKALINE PHOSPHATASE from Biocatalysts

Biocatalysts Limited, the British speciality enzyme company, has developed a completely new type of alkaline phosphatase with many advantages over the types most commonly used.

It is of microbial origin with a high specific activity (unlike that from *E. coli*) and with higher temperature and storage stability compared to that from calf intestine.

This is the first of several new generation diagnostic enzymes being developed by Biocatalysts Limited with greatly improved stability.

- Non-animal source, no risk of BSE or animal virus contamination
- Higher temperature stability than calf intestine
- Much higher specific activity than from *E. coli*
- Very high storage stability even in the absence of glycerol

For further details on alkaline phosphatase and our other diagnostic enzymes contact us direct at the address below or within North America contact our US Distributor Kaltron-Pettibone 'phone: 630 350 1116 or fax: 630-350-1606

Biocatalysts Limited
Treforest Industrial Estate Pontypridd Wales UK CF37 5UD
Tel: +44 (0)1443 843712 Fax: +44 (0)1443 841214
e-mail: Kelly@Biocatalysts.com.



- Fischer-Vize, *Science* 270, 1828 (1995).
35. T. C. James and S. C. Elgin, *Mol. Cell Biol.* 6, 3862 (1986); R. Paro and D. S. Hogness, *Proc. Natl. Acad. Sci. U.S.A.* 88, 263 (1991); B. Tschiersch et al., *EMBO J.* 13, 3822 (1994); M. T. Madireddi et al., *Cell* 87, 75 (1996); D. G. Stokes, K. D. Tartof, R. P. Perry, *Proc. Natl. Acad. Sci. U.S.A.* 93, 7137 (1996).
36. P. M. Palosaari et al., *J. Biol. Chem.* 266, 10750 (1991); A. Schmitz, K. H. Gartemann, J. Fiedler, E.

- Grund, R. Eichenlaub, *Appl. Environ. Microbiol.* 58, 4068 (1992); V. Sharma, K. Suvama, R. Meganathan, M. E. Hudspeth, *J. Bacteriol.* 174, 5057 (1992); M. Kanazawa et al., *Enzyme Protein* 47, 9 (1993); Z. L. Boynton, G. N. Bennet, F. B. Rudolph, *J. Bacteriol.* 178, 3015 (1996).
37. M. Ho et al., *Cell* 77, 869 (1994).
38. W. Hendriks et al., *J. Cell Biochem.* 59, 418 (1995).
39. We thank H. Skaletsky and F. Lewitter for help with

sequence analysis; Lawrence Livermore National Laboratory for the flow-sorted Y cosmid library; and P. Bain, A. Bortvin, A. de la Chapelle, G. Fink, K. Jegalian, T. Kawaguchi, E. Lander, H. Lodish, P. Matsudaira, D. Menke, U. RajBhandary, R. Reijo, S. Rozen, A. Schwartz, C. Sun, and C. Tilford for comments on the manuscript. Supported by NIH.

28 April 1997; accepted 9 September 1997

Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale

Joseph L. DeRisi, Vishwanath R. Iyer, Patrick O. Brown*

DNA microarrays containing virtually every gene of *Saccharomyces cerevisiae* were used to carry out a comprehensive investigation of the temporal program of gene expression accompanying the metabolic shift from fermentation to respiration. The expression profiles observed for genes with known metabolic functions pointed to features of the metabolic reprogramming that occur during the diauxic shift, and the expression patterns of many previously uncharacterized genes provided clues to their possible functions. The same DNA microarrays were also used to identify genes whose expression was affected by deletion of the transcriptional co-repressor *TUP1* or overexpression of the transcriptional activator *YAP1*. These results demonstrate the feasibility and utility of this approach to genomewide exploration of gene expression patterns.

The complete sequences of nearly a dozen microbial genomes are known, and in the next several years we expect to know the complete genome sequences of several metazoans, including the human genome. Defining the role of each gene in these genomes will be a formidable task, and understanding how the genome functions as a whole in the complex natural history of a living organism presents an even greater challenge.

Knowing when and where a gene is expressed often provides a strong clue as to its biological role. Conversely, the pattern of genes expressed in a cell can provide detailed information about its state. Although regulation of protein abundance in a cell is by no means accomplished solely by regulation of mRNA, virtually all differences in cell type or state are correlated with changes in the mRNA levels of many genes. This is fortuitous because the only specific reagent required to measure the abundance of the mRNA for a specific gene is a cDNA sequence. DNA microarrays, consisting of thousands of individual gene sequences printed in a high-density array on a glass microscope slide (1, 2), provide a practical and economical tool for studying gene expression on a very large scale (3–6).

Saccharomyces cerevisiae is an especially

favorable organism in which to conduct a systematic investigation of gene expression. The genes are easy to recognize in the genome sequence, cis regulatory elements are generally compact and close to the transcription units, much is already known about its genetic regulatory mechanisms, and a powerful set of tools is available for its analysis.

A recurring cycle in the natural history of yeast involves a shift from anaerobic (fermentation) to aerobic (respiration) metabolism. Inoculation of yeast into a medium rich in sugar is followed by rapid growth fueled by fermentation, with the production of ethanol. When the fermentable sugar is exhausted, the yeast cells turn to ethanol as a carbon source for aerobic growth. This switch from anaerobic growth to aerobic respiration upon depletion of glucose, referred to as the diauxic shift, is correlated with widespread changes in the expression of genes involved in fundamental cellular processes such as carbon metabolism, protein synthesis, and carbohydrate storage (7). We used DNA microarrays to characterize the changes in gene expression that take place during this process for nearly the entire genome, and to investigate the genetic circuitry that regulates and executes this program.

Yeast open reading frames (ORFs) were amplified by the polymerase chain reaction (PCR), with a commercially available set of primer pairs (8). DNA microarrays, containing approximately 6400 distinct DNA sequences, were printed onto glass slides by

using a simple robotic printing device (9). Cells from an exponentially growing culture of yeast were inoculated into fresh medium and grown at 30°C for 21 hours. After an initial 9 hours of growth, samples were harvested at seven successive 2-hour intervals, and mRNA was isolated (10). Fluorescently labeled cDNA was prepared by reverse transcription in the presence of Cy3(green)- or Cy5(red)-labeled deoxyuridine triphosphate (dUTP) (11) and then hybridized to the microarrays (12). To maximize the reliability with which changes in expression levels could be discerned, we labeled cDNA prepared from cells at each successive time point with Cy5, then mixed it with a Cy3-labeled "reference" cDNA sample prepared from cells harvested at the first interval after inoculation. In this experimental design, the relative fluorescence intensity measured for the Cy3 and Cy5 fluors at each array element provides a reliable measure of the relative abundance of the corresponding mRNA in the two cell populations (Fig. 1). Data from the series of seven samples (Fig. 2), consisting of more than 43,000 expression-ratio measurements, were organized into a database to facilitate efficient exploration and analysis of the results. This database is publicly available on the Internet (13).

During exponential growth in glucose-rich medium, the global pattern of gene expression was remarkably stable. Indeed, when gene expression patterns between the first two cell samples (harvested at a 2-hour interval) were compared, mRNA levels differed by a factor of 2 or more for only 19 genes (0.3%), and the largest of these differences was only 2.7-fold (14). However, as glucose was progressively depleted from the growth media during the course of the experiment, a marked change was seen in the global pattern of gene expression. mRNA levels for approximately 710 genes were induced by a factor of at least 2, and the mRNA levels for approximately 1030 genes declined by a factor of at least 2. Messenger RNA levels for 183 genes increased by a factor of at least 4, and mRNA levels for 203 genes diminished by a factor of at least 4. About half of these differentially expressed genes have no currently recognized function and are not yet named. Indeed, more than 400 of the differentially expressed genes have no apparent homology

Department of Biochemistry, Stanford University School of Medicine, Howard Hughes Medical Institute, Stanford, CA 94305–5428, USA.

*To whom correspondence should be addressed. E-mail: pbrown@cmgm.stanford.edu

to any gene whose function is known (15). The responses of these previously uncharacterized genes to the diauxic shift therefore provides the first small clue to their possible roles.

The global view of changes in expression of genes with known functions provides a vivid picture of the way in which the cell adapts to a changing environment. Figure 3 shows a portion of the yeast metabolic pathways involved in carbon and energy metabolism. Mapping the changes we observed in the mRNAs encoding each enzyme onto this framework allowed us to infer the redirection in the flow of metabolites through this system. We observed large inductions of the genes coding for the enzymes aldehyde dehydrogenase (ALD2) and acetyl-coenzyme A (CoA) synthase (ACS1), which function together to convert the products of alcohol dehydrogenase into acetyl-CoA, which in turn is used to fuel the tricarboxylic acid (TCA) cycle and the glyoxylate cycle. The concomitant shutdown of transcription of the genes encoding pyruvate decarboxylase and induction of pyruvate carboxylase rechannels pyruvate away from acetaldehyde, and instead to oxalacetate, where it can serve to supply the TCA cycle and gluconeogenesis. Induction of the pivotal genes *PCK1*, encoding phosphoenolpyruvate carboxykinase, and *FBP1*, encoding fructose 1,6-bisphosphatase, switches the directions of two key irreversible steps in glycolysis, reversing the flow of metabolites along the reversible steps of the glycolytic pathway toward the essential biosynthetic precursor, glucose-6-phosphate. Induction of the genes coding for the trehalose synthase and glycogen synthase complexes promotes channeling of glucose-6-phosphate into these carbohydrate storage pathways.

Just as the changes in expression of genes encoding pivotal enzymes can provide insight into metabolic reprogramming, the behavior of large groups of functionally related genes can provide a broad view of the systematic way in which the yeast cell adapts to a changing environment (Fig. 4). Several classes of genes, such as cytochrome c-related genes and those involved in the TCA/glyoxylate cycle and carbohydrate storage, were coordinately induced by glucose exhaustion. In contrast, genes devoted to protein synthesis, including ribosomal proteins, tRNA synthetases, and translation, elongation, and initiation factors, exhibited a coordinated decrease in expression. More than 95% of ribosomal genes showed at least twofold decreases in expression during the diauxic shift (Fig. 4) (13). A noteworthy and illuminating exception was that the

genes encoding mitochondrial ribosomal genes were generally induced rather than repressed after glucose limitation, highlighting the requirement for mitochondrial biogenesis (13). As more is learned about the functions of every gene in the yeast genome, the ability to gain insight into a cell's response to a changing environment through its global gene expression patterns will become increasingly powerful.

Several distinct temporal patterns of expression could be recognized, and sets of genes could be grouped on the basis of the similarities in their expression patterns. The characterized members of each of these groups also shared important similarities in their functions. Moreover, in most cases, common regulatory mechanisms could be inferred for sets of genes with similar expression profiles. For example, seven genes showed a late induction profile, with mRNA levels increasing by more than ninefold at

the last timepoint but less than threefold at the preceding timepoint (Fig. 5B). All of these genes were known to be glucose-repressed, and five of the seven were previously noted to share a common upstream activating sequence (UAS), the carbon source response element (CSRE) (16–20). A search in the promoter regions of the remaining two genes, *ACR1* and *IDP2*, revealed that *ACR1*, a gene essential for *ACS1* activity, also possessed a consensus CSRE motif, but interestingly, *IDP2* did not. A search of the entire yeast genome sequence for the consensus CSRE motif revealed only four additional candidate genes, none of which showed a similar induction.

Examples from additional groups of genes that shared expression profiles are illustrated in Fig. 5, C through F. The sequences upstream of the named genes in Fig. 5C all contain stress response elements (STRE), and with the exception

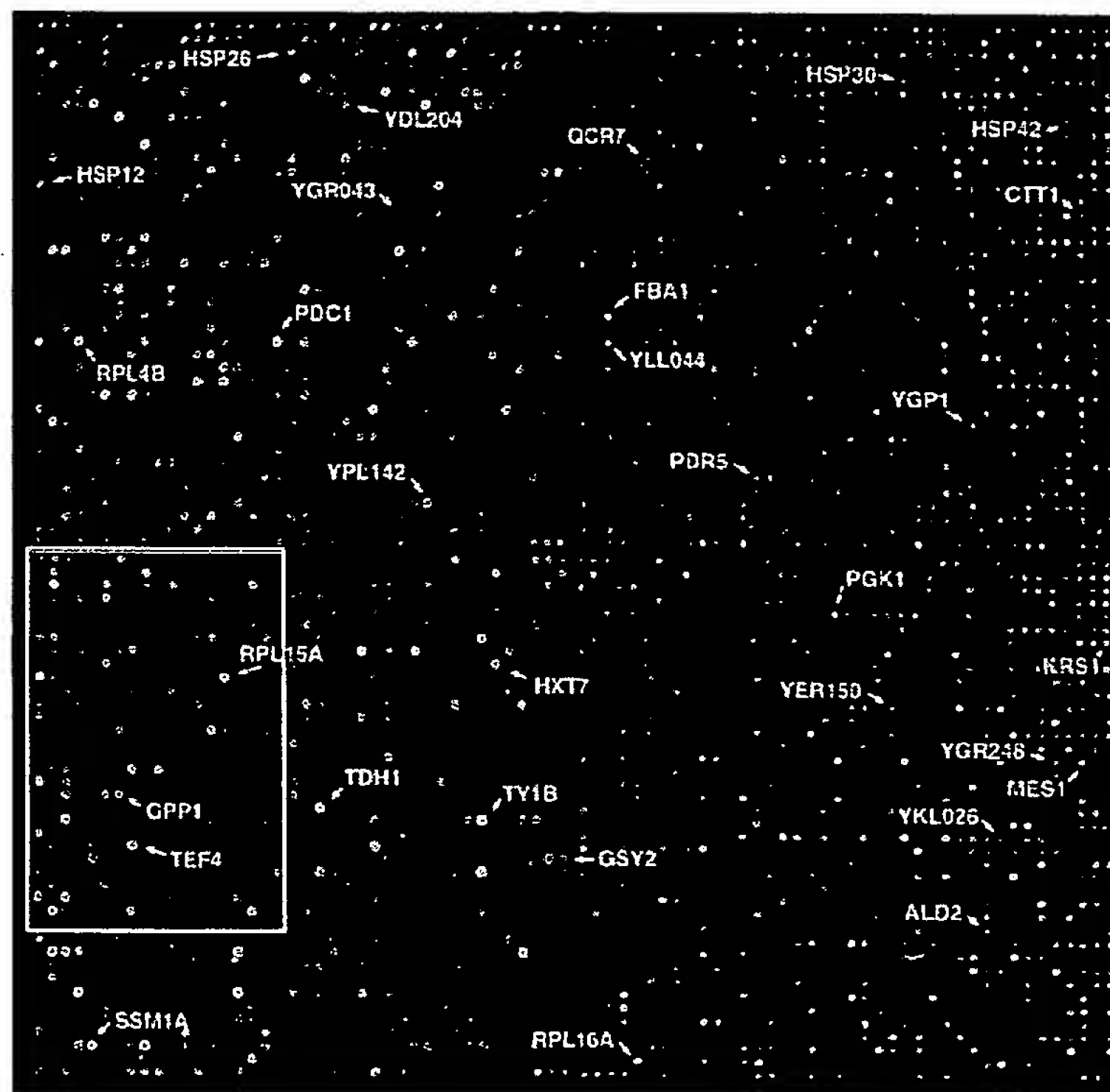


Fig. 1. Yeast genome microarray. The actual size of the microarray is 18 mm by 18 mm. The microarray was printed as described (9). This image was obtained with the same fluorescent scanning confocal microscope used to collect all the data we report (49). A fluorescently labeled cDNA probe was prepared from mRNA isolated from cells harvested shortly after inoculation (culture density of $<5 \times 10^6$ cells/ml and media glucose level of 19 g/liter) by reverse transcription in the presence of Cy3-dUTP. Similarly, a second probe was prepared from mRNA isolated from cells taken from the same culture 9.5 hours later (culture density of $\sim 2 \times 10^8$ cells/ml, with a glucose level of <0.2 g/liter) by reverse transcription in the presence of Cy5-dUTP. In this image, hybridization of the Cy3-dUTP-labeled cDNA (that is, mRNA expression at the initial timepoint) is represented as a green signal, and hybridization of Cy5-dUTP-labeled cDNA (that is, mRNA expression at 9.5 hours) is represented as a red signal. Thus, genes induced or repressed after the diauxic shift appear in this image as red and green spots, respectively. Genes expressed at roughly equal levels before and after the diauxic shift appear in this image as yellow spots.

of *HSP42*, have previously been shown to be controlled at least in part by these elements (21–24). Inspection of the sequences upstream of *HSP42* and the two uncharacterized genes shown in Fig. 5C, YKL026c, a hypothetical protein with similarity to glutathione peroxidase, and YGR043c, a putative transaldolase, revealed that each of these genes also possess repeated upstream copies of the stress-responsive CCCCT motif. Of the 13 additional genes in the yeast genome that shared this expression profile [including *HSP30*, *ALD2*, *OM45*, and 10 uncharacterized ORFs (25)], nine contained one or more recognizable STRE sites in their upstream regions.

The heterotrimeric transcriptional activator complex HAP2,3,4 has been shown to be responsible for induction of several genes important for respiration (26–28). This complex binds a degenerate consensus sequence known as the CCAAT box (26). Computer analysis, using the consensus sequence TNRYTGGB (29), has suggested that a large number of genes involved in respiration may be specific targets of HAP2,3,4 (30). Indeed, a putative HAP2,3,4 binding site could be found in the sequences upstream of each of the seven cytochrome *c*-related genes that showed the greatest magnitude of induction (Fig. 5D). Of 12 additional cytochrome *c*-related genes that were induced, HAP2,3,4 binding sites were present in all but one. Significantly, we found that transcription of *HAP4* itself was induced nearly ninefold concomitant with the diauxic shift.

Control of ribosomal protein biogenesis is mainly exerted at the transcriptional level, through the presence of a common upstream-activating element (UAS_{rp}) that is recognized by the Rap1 DNA-binding protein (31, 32). The expression profiles of seven ribosomal proteins are shown in Fig. 5F. A search of the sequences upstream of all seven genes revealed consensus Rap1-binding motifs (33). It has been suggested that declining Rap1 levels in the cell during starvation may be responsible for the decline in ribosomal protein gene expression (34). Indeed, we observed that the abundance of RAP1 mRNA diminished by 4.4-fold, at about the time of glucose exhaustion.

Of the 149 genes that encode known or putative transcription factors, only two, *HAP4* and *SIP4*, were induced by a factor of more than threefold at the diauxic shift. *SIP4* encodes a DNA-binding transcriptional activator that has been shown to interact with Snf1, the “master regulator” of glucose repression (35). The eightfold induction of *SIP4* upon depletion of glucose strongly suggests a role in the induction of

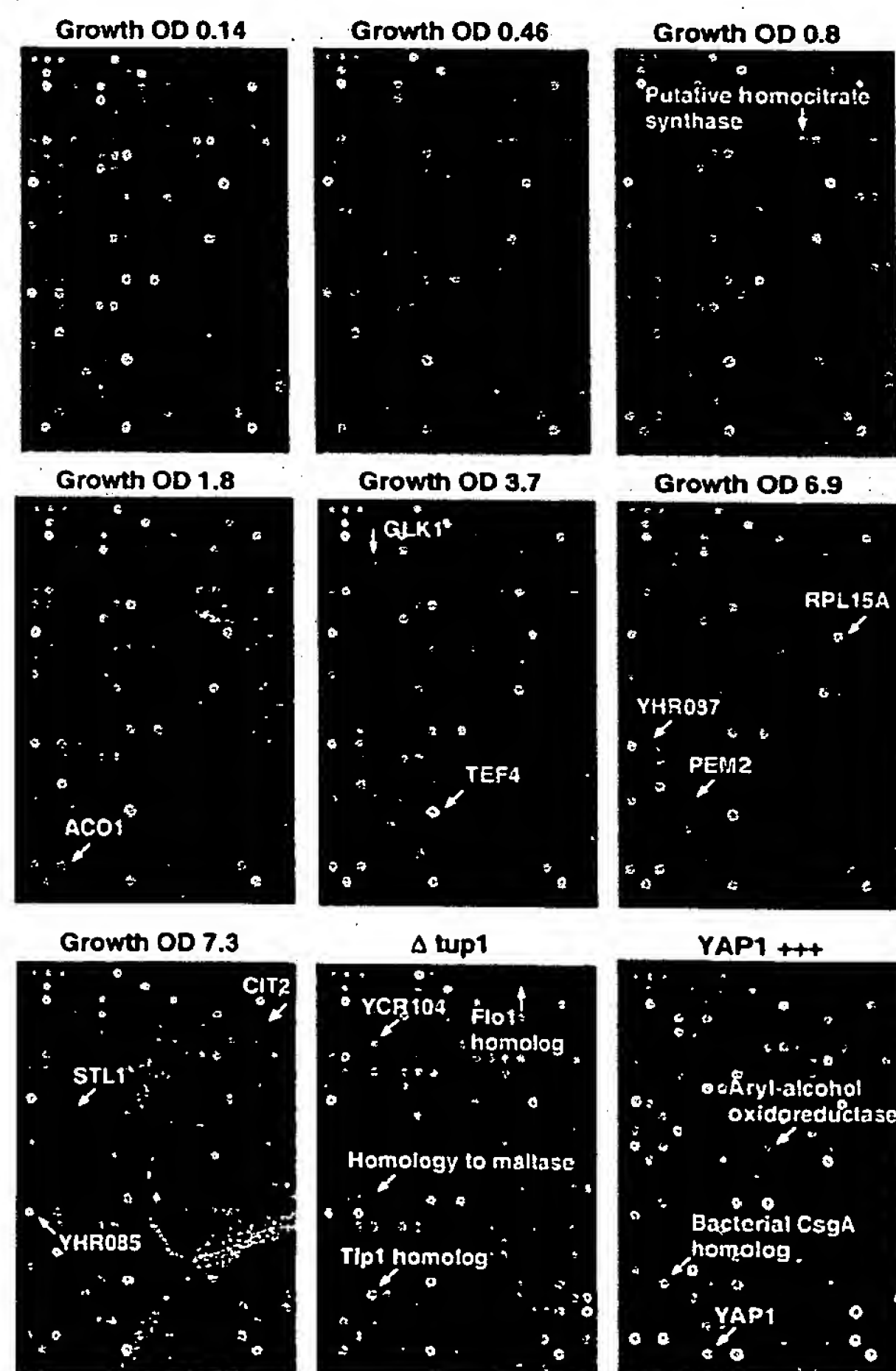
downstream genes at the diauxic shift.

Although most of the transcriptional responses that we observed were not previously known, the responses of many genes during the diauxic shift have been described. Comparison of the results we obtained by DNA microarray hybridization with previously reported results therefore provided a strong test of the sensitivity and accuracy of this approach. The expression patterns we observed for previously characterized genes showed almost perfect concordance with previously published results (36). Moreover, the differential expression measurements obtained by DNA microarray hybridization were reproducible in duplicate experiments. For example, the remarkable changes in gene expression between cells harvested immediately after inoculation and immediately after the diauxic shift (the first and sixth intervals in this time series) were measured in duplicate, independent DNA microarray hybridizations. The correlation coefficient for two complete sets of expression ratio measurements was 0.87, and for more than 95% of the genes, the expres-

sion ratios measured in these duplicate experiments differed by less than a factor of 2. However, in a few cases, there were discrepancies between our results and previous results, pointing to technical limitations that will need to be addressed as DNA microarray technology advances (37, 38). Despite the noted exceptions, the high concordance between the results we obtained in these experiments and those of previous studies provides confidence in the reliability and thoroughness of the survey.

The changes in gene expression during this diauxic shift are complex and involve integration of many kinds of information about the nutritional and metabolic state of the cell. The large number of genes whose expression is altered and the diversity of temporal expression profiles observed in this experiment highlight the challenge of understanding the underlying regulatory mechanisms. One approach to defining the contributions of individual regulatory genes to a complex program of this kind is to use DNA microarrays to identify genes whose expression is affected

Fig. 2. The section of the array indicated by the gray box in Fig. 1 is shown for each of the experiments described here. Representative genes are labeled. In each of the arrays used to analyze gene expression during the diauxic shift, red spots represent genes that were induced relative to the initial timepoint, and green spots represent genes that were repressed relative to the initial timepoint. In the arrays used to analyze the effects of the *tup1Δ* mutation and *YAP1* overexpression, red spots represent genes whose expression was increased, and green spots represent genes whose expression was decreased by the genetic modification. Note that distinct sets of genes are induced and repressed in the different experiments. The complete images of each of these arrays can be viewed on the Internet (13). Cell density as measured by optical density (OD) at 600 nm was used to measure the growth of the culture.



by mutations in each putative regulatory gene. As a test of this strategy, we analyzed the genomewide changes in gene expression that result from deletion of the *TUP1* gene. Transcriptional repression of many genes by glucose requires the DNA-binding repressor

Mig1 and is mediated by recruiting the transcriptional co-repressors Tup1 and Cyc8/Ssn6 (39). Tup1 has also been implicated in repression of oxygen-regulated, mating-type-specific, and DNA-damage-inducible genes (40).

Wild-type yeast cells and cells bearing a deletion of the *TUP1* gene (*tup1Δ*) were grown in parallel cultures in rich medium containing glucose as the carbon source. Messenger RNA was isolated from exponentially growing cells from the two populations and used to prepare cDNA labeled with Cy3 (green) and Cy5 (red), respectively (11). The labeled probes were mixed and simultaneously hybridized to the microarray. Red spots on the microarray therefore represented genes whose transcription was induced in the *tup1Δ* strain, and thus presumably repressed by Tup1 (41). A representative section of the microarray (Fig. 2, bottom middle panel) illustrates that the genes whose expression was affected by the *tup1Δ* mutation, were, in general, distinct from those induced upon glucose exhaustion [complete images of all the arrays shown in Fig. 2 are available on the Internet (13)]. Nevertheless, 34 (10%) of the genes that were induced by a factor of at least 2 after the diauxic shift were similarly induced by deletion of *TUP1*, suggesting that these genes may be subject to *TUP1*-mediated repression by glucose. For example, *SUC2*, the gene encoding invertase, and all five hexose transporter genes that were induced during the course of the diauxic shift were similarly induced, in duplicate experiments, by the deletion of *TUP1*.

The set of genes affected by Tup1 in this experiment also included α -glucosidases, the mating-type-specific genes *MFA1* and *MFA2*, and the DNA damage-inducible *RNR2* and *RNR4*, as well as genes involved in flocculation and many genes of unknown function. The hybridization signal corresponding to expression of *TUP1* itself was also severely reduced because of the (incomplete) deletion of the transcription unit in the *tup1Δ* strain, providing a positive control in the experiment (42).

Many of the transcriptional targets of Tup1 fell into sets of genes with related biochemical functions. For instance, although only about 3% of all yeast genes appeared to be *TUP1*-repressed by a factor of more than 2 in duplicate experiments under these conditions, 6 of the 13 genes that have been implicated in flocculation (15) showed a reproducible increase in expression of at least twofold when *TUP1* was deleted. Another group of related genes that appeared to be subject to *TUP1* repression encodes the serine-rich cell wall mannoproteins, such as *Tip1* and *Tir1/Srp1* which are induced by cold shock and other stresses (43), and similar, serine-poor proteins, the seripauperins (44). Messenger RNA levels for 23 of the 26 genes in this group were reproducibly elevated by at least 2.5-fold in the *tup1Δ*

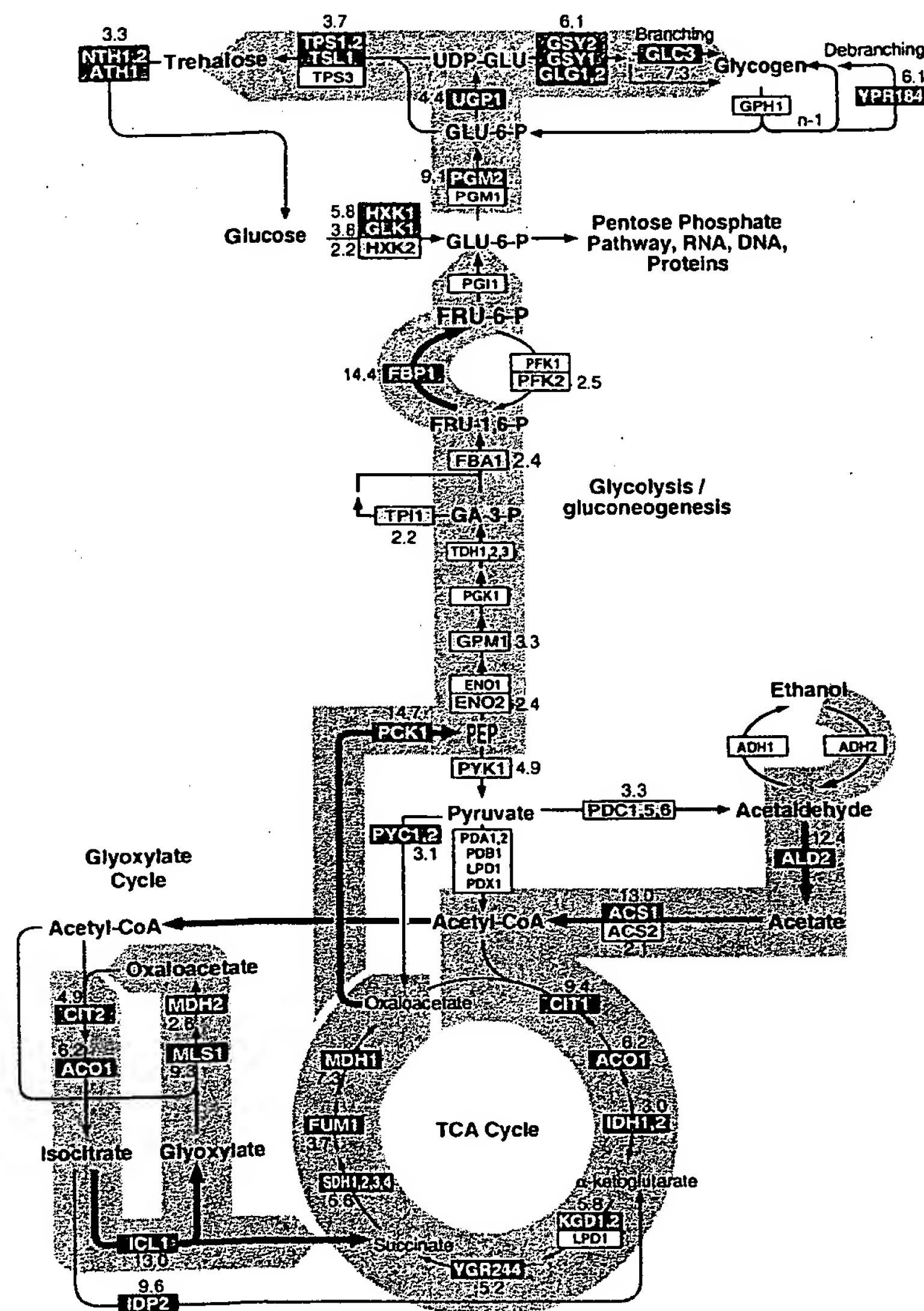


Fig. 3. Metabolic reprogramming inferred from global analysis of changes in gene expression. Only key metabolic intermediates are identified. The yeast genes encoding the enzymes that catalyze each step in this metabolic circuit are identified by name in the boxes. The genes encoding succinyl-CoA synthase and glycogen-debranching enzyme have not been explicitly identified, but the ORFs YGR244 and YPR184 show significant homology to known succinyl-CoA synthase and glycogen-debranching enzymes, respectively, and are therefore included in the corresponding steps in this figure. Red boxes with white lettering identify genes whose expression increases in the diauxic shift. Green boxes with dark green lettering identify genes whose expression diminishes in the diauxic shift. The magnitude of induction or repression is indicated for these genes. For multimeric enzyme complexes, such as succinate dehydrogenase, the indicated fold-induction represents an unweighted average of all the genes listed in the box. Black and white boxes indicate no significant differential expression (less than twofold). The direction of the arrows connecting reversible enzymatic steps indicate the direction of the flow of metabolic intermediates, inferred from the gene expression pattern, after the diauxic shift. Arrows representing steps catalyzed by genes whose expression was strongly induced are highlighted in red. The broad gray arrows represent major increases in the flow of metabolites after the diauxic shift, inferred from the indicated changes in gene expression.

strain, and 18 of these genes were induced by more than sevenfold when *TUP1* was deleted. In contrast, none of 83 genes that could be classified as putative regulators of the cell division cycle were induced more than twofold by deletion of *TUP1*. Thus, despite the diversity of the regulatory systems that employ Tup1, most of the genes that it regulates under these conditions fall into a limited number of distinct functional classes.

Because the microarray allows us to monitor expression of nearly every gene in yeast, we can, in principle, use this approach to identify all the transcriptional targets of a regulatory protein like Tup1. It is important to note, however, that in any single experiment of this kind we can only recognize those target genes that are normally repressed (or induced) under the conditions of the experiment. For instance, the experiment described here analyzed a MAT α strain in which *MFA1* and *MFA2*, the genes encoding the α -factor mating pheromone precursor, are normally repressed. In the isogenic *tup1* Δ strain, these genes were inappropriately expressed, reflecting the role that Tup1 plays in their repression. Had we instead carried out this experiment with a MAT α strain (in which expression of *MFA1* and *MFA2* is not repressed), it would not have been possible to conclude anything regarding the role of Tup1 in the repression of these genes. Conversely, we cannot distinguish indirect effects of the chronic absence of Tup1 in the mutant strain from effects directly attributable to its participation in repressing the transcription of a gene.

Another simple route to modulating the activity of a regulatory factor is to overexpress the gene that encodes it. *YAP1* encodes a DNA-binding transcription factor belonging to the bZIP class of DNA-binding proteins. Overexpression of *YAP1* in yeast confers increased resistance to hydrogen peroxide, *o*-phenanthroline, heavy metals, and osmotic stress (45). We analyzed differential gene expression between a wild-type strain bearing a control plasmid and a strain with a plasmid expressing *YAP1* under the control of the strong *GALI-10* promoter, both grown in galactose (that is, a condition that induces *YAP1* overexpression). Complementary DNA from the control and *YAP1* overexpressing strains, labeled with Cy3 and Cy5, respectively, was prepared from mRNA isolated from the two strains and hybridized to the microarray. Thus, red spots on the array represent genes that were induced in the strain overexpressing *YAP1*.

Of the 17 genes whose mRNA levels increased by more than threefold when

YAP1 was overexpressed in this way, five bear homology to aryl-alcohol oxidoreductases (Fig. 2 and Table 1). An additional four of the genes in this set also belong to the general class of dehydrogenases/oxidoreductases. Very little is known about the role of aryl-alcohol oxidoreductases in *S. cerevisiae*, but these enzymes have been isolated from ligninolytic fungi, in which they participate in coupled redox reactions, oxidizing aromatic, and aliphatic unsaturated alcohols to aldehydes with the production of hydrogen peroxide (46, 47). The fact that a remarkable fraction of the targets identified in this experiment belong to the same small, functional group of oxidoreductases suggests that these genes

might play an important protective role during oxidative stress. Transcription of a small number of genes was reduced in the strain overexpressing *Yap1*. Interestingly, many of these genes encode sugar permeases or enzymes involved in inositol metabolism.

We searched for *Yap1*-binding sites (TTACTAA or TGACTAA) in the sequences upstream of the target genes we identified (48). About two-thirds of the genes that were induced by more than threefold upon *Yap1* overexpression had one or more binding sites within 600 bases upstream of the start codon (Table 1), suggesting that they are directly regulated by *Yap1*. The absence of canonical *Yap1*-bind-

Fig. 4. Coordinated regulation of functionally related genes. The curves represent the average induction or repression ratios for all the genes in each indicated group. The total number of genes in each group was as follows: ribosomal proteins, 112; translation elongation and initiation factors, 25; tRNA synthetases (excluding mitochondrial synthetases), 17; glycogen and trehalose synthesis and degradation, 15; cytochrome c oxidase and reductase proteins, 19; and TCA- and glyoxylate-cycle enzymes, 24.

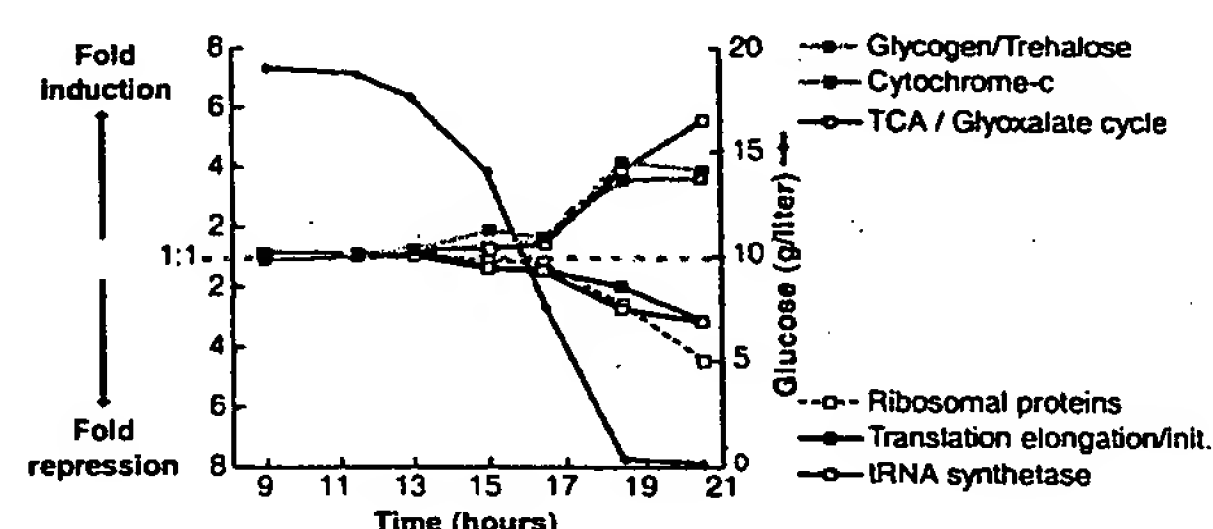


Table 1. Genes induced by *YAP1* overexpression. This list includes all the genes for which mRNA levels increased by more than twofold upon *YAP1* overexpression in both of two duplicate experiments, and for which the average increase in mRNA level in the two experiments was greater than threefold (50). Positions of the canonical *Yap1* binding sites upstream of the start codon, when present, and the average fold-increase in mRNA levels measured in the two experiments are indicated.

ORF	Distance of <i>Yap1</i> site from ATG	Gene	Description	Fold-increase
YNL331C	162–222 (5 sites)	<i>YAP1</i>	Putative aryl-alcohol reductase	12.9
YKL071W			Similarity to bacterial <i>csgA</i> protein	10.4
YML007W			Transcriptional activator involved in oxidative stress response	9.8
YFL056C	223, 242		Homology to aryl-alcohol dehydrogenases	9.0
YLL060C	98		Putative glutathione transferase	7.4
YOL165C	266		Putative aryl-alcohol dehydrogenase (NADP+)	7.0
YCR107W	409	<i>ATR1</i>	Putative aryl-alcohol reductase	6.5
YML116W			Aminotriazole and 4-nitroquinoline resistance protein	6.5
YBR008C	142, 167, 364		Homology to benomyl/methotrexate resistance protein	6.1
YCLX08C	148, 212	<i>OYE3</i>	Hypothetical protein	6.1
YJR155W			Putative aryl-alcohol dehydrogenase	6.0
YPL171C			NAPDH dehydrogenase (old yellow enzyme), isoform 3	5.8
YLR460C	167, 317		Homology to hypothetical proteins YCR102c and YNL134c	4.7
YKR076W	178		Homology to hypothetical protein YMR251w	4.5
YHR179W	327	<i>OYE2</i>	NAD(P)H oxidoreductase (old yellow enzyme), isoform 1	4.1
YML131W	507		Similarity to <i>A. thaliana</i> zeta-crystallin homolog	3.7
YOL126C		<i>MDH2</i>	Malate dehydrogenase	3.3

ing sites upstream of the others may reflect an ability of Yap1 to bind sites that differ from the canonical binding sites, perhaps in cooperation with other factors, or less likely, may represent an indirect effect of Yap1 overexpression, mediated by one or more intermediary factors. Yap1 sites were found only four times in the corresponding region of an arbitrary set of 30 genes that were not differentially regulated by Yap1.

Use of a DNA microarray to characterize the transcriptional consequences of mutations affecting the activity of regulatory molecules provides a simple and powerful approach to dissection and characterization of regulatory pathways and net-

works. This strategy also has an important practical application in drug screening. Mutations in specific genes encoding candidate drug targets can serve as surrogates for the ideal chemical inhibitor or modulator of their activity. DNA microarrays can be used to define the resulting signature pattern of alterations in gene expression, and then subsequently used in an assay to screen for compounds that reproduce the desired signature pattern.

DNA microarrays provide a simple and economical way to explore gene expression patterns on a genomic scale. The hurdles to extending this approach to any other organism are minor. The equipment

required for fabricating and using DNA microarrays (9) consists of components that were chosen for their modest cost and simplicity. It was feasible for a small group to accomplish the amplification of more than 6000 genes in about 4 months and, once the amplified gene sequences were in hand, only 2 days were required to print a set of 110 microarrays of 6400 elements each. Probe preparation, hybridization, and fluorescent imaging are also simple procedures. Even conceptually simple experiments, as we described here, can yield vast amounts of information. The value of the information from each experiment of this kind will progressively increase as more is learned about the functions of each gene and as additional experiments define the global changes in gene expression in diverse other natural processes and genetic perturbations. Perhaps the greatest challenge now is to develop efficient methods for organizing, distributing, interpreting, and extracting insights from the large volumes of data these experiments will provide.

REFERENCES AND NOTES

1. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, *Science* 270, 467 (1995).
2. D. Shalon, S. J. Smith, P. O. Brown, *Genome Res.* 6, 639 (1996).
3. D. Lashkari, *Proc. Natl. Acad. Sci. U.S.A.*, in press.
4. J. DeRisi et al., *Nature Genet.* 14, 457 (1996).
5. D. J. Lockhart et al., *Nature Biotechnol.* 14, 1675 (1996).
6. M. Chee et al., *Science* 274, 610 (1996).
7. M. Johnston and M. Carlson, in *The Molecular Biology of the Yeast Saccharomyces: Gene Expression*, E. W. Jones, J. R. Pringle, J. R. Broach, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1992), p. 193.
8. Primers for each known or predicted protein coding sequence were supplied by Research Genetics. PCR was performed with the protocol supplied by Research Genetics, using genomic DNA from yeast strain S288C as a template. Each PCR product was verified by agarose gel electrophoresis and was deemed correct if the lane contained a single band of appropriate mobility. Failures were marked as such in the database. The overall success rate for a single-pass amplification of 6116 ORFs was ~94.5%.
9. Glass slides (Gold Seal) were cleaned for 2 hours in a solution of 2 N NaOH and 70% ethanol. After rinsing in distilled water, the slides were then treated with a 1:5 dilution of poly-L-lysine adhesive solution (Sigma) for 1 hour, and then dried for 5 min at 40°C in a vacuum oven. DNA samples from 100- μ l PCR reactions were purified by ethanol purification in 96-well microtiter plates. The resulting precipitates were resuspended in 3 \times standard saline citrate (SSC) and transferred to new plates for arraying. A custom-built arraying robot was used to print on a batch of 110 slides. Details of the design of the microarrayer are available at cmgm.stanford.edu/pbrown. After printing, the microarrays were rehydrated for 30 s in a humid chamber and then snap-dried for 2 s on a hot plate (100°C). The DNA was then ultraviolet (UV)-crosslinked to the surface by subjecting the slides to 60 mJ of energy (Stratagene Stratalinker). The rest of the poly-L-lysine surface was blocked by a 15-min incubation in a solution of 70 mM succinic anhydride dissolved in a solution consisting of 315 ml of 1-methyl-2-pyrrolidinone (Aldrich) and 35 ml of 1 M boric acid (pH 8.0). Directly after the blocking reac-

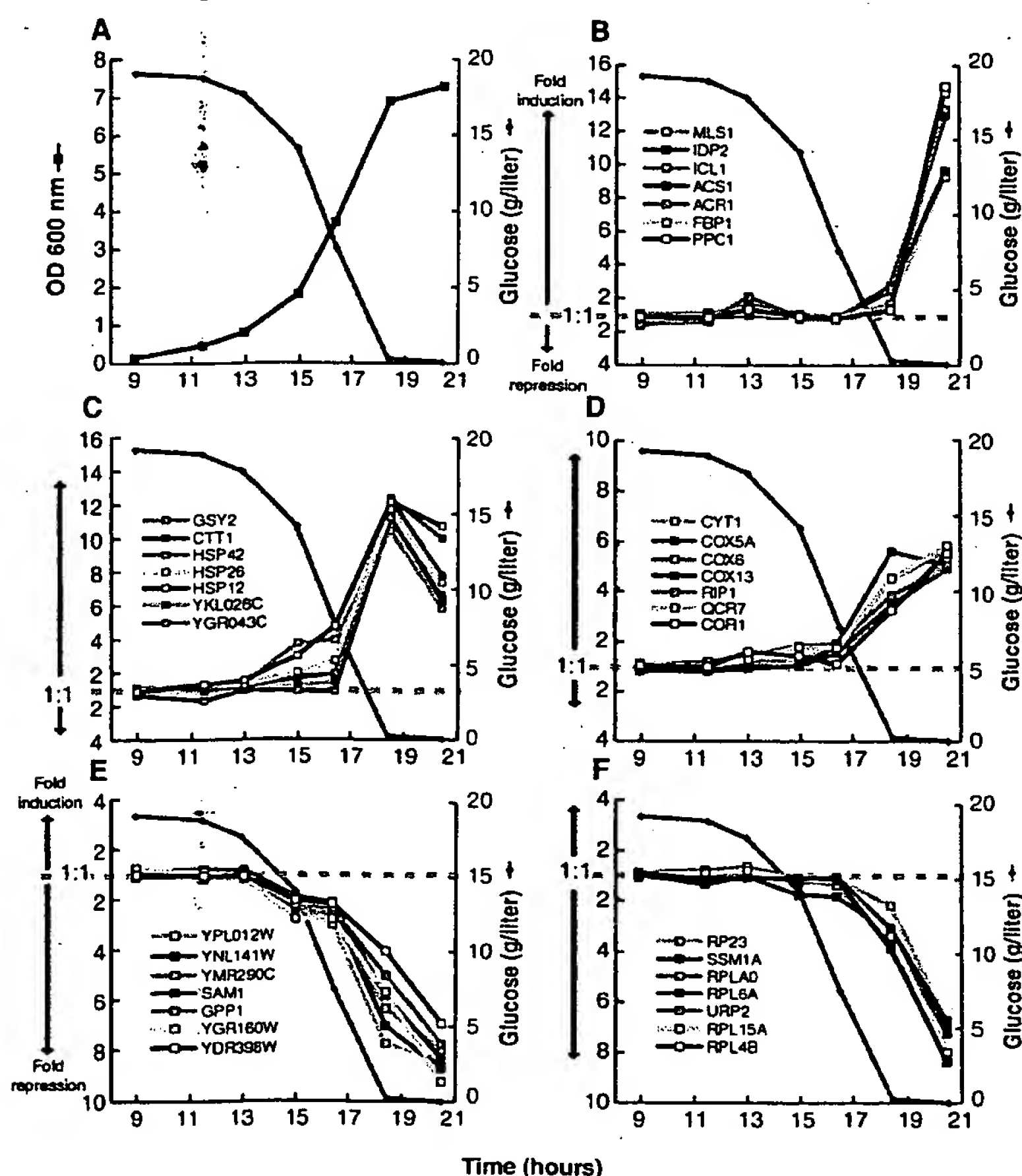


Fig. 5. Distinct temporal patterns of induction or repression help to group genes that share regulatory properties. (A) Temporal profile of the cell density, as measured by OD at 600 nm and glucose concentration in the media. (B) Seven genes exhibited a strong induction (greater than ninefold) only at the last timepoint (20.5 hours). With the exception of *IDP2*, each of these genes has a CSRE UAS. There were no additional genes observed to match this profile. (C) Seven members of a class of genes marked by early induction with a peak in mRNA levels at 18.5 hours. Each of these genes contains STRE motif repeats in their upstream promoter regions. (D) Cytochrome c oxidase and ubiquinol cytochrome c reductase genes. Marked by an induction coincident with the diauxic shift, each of these genes contains a consensus binding motif for the HAP2,3,4 protein complex. At least 17 genes shared a similar expression profile. (E) *SAM1*, *GPP1*, and several genes of unknown function are repressed before the diauxic shift, and continue to be repressed upon entry into stationary phase. (F) Ribosomal protein genes comprise a large class of genes that are repressed upon depletion of glucose. Each of the genes profiled here contains one or more RAP1-binding motifs upstream of its promoter. RAP1 is a transcriptional regulator of most ribosomal proteins.

- tion, the bound DNA was denatured by a 2-min incubation in distilled water at ~95°C. The slides were then transferred into a bath of 100% ethanol at room temperature, rinsed, and then spun dry in a clinical centrifuge. Slides were stored in a closed box at room temperature until used.
10. YPD medium (8 liters), in a 10-liter fermentation vessel, was inoculated with 2 ml of a fresh overnight culture of yeast strain DBY7286 (MATa, ura3, GAL2). The fermentor was maintained at 30°C with constant agitation and aeration. The glucose content of the media was measured with a UV test kit (Boehringer Mannheim, catalog number 716251). Cell density was measured by OD at 600-nm wavelength. Aliquots of culture were rapidly withdrawn from the fermentation vessel by peristaltic pump, spun down at room temperature, and then flash frozen with liquid nitrogen. Frozen cells were stored at -80°C.
 11. Cy3-dUTP or Cy5-dUTP (Amersham) was incorporated during reverse transcription of 1.25 µg of polyadenylated [poly(A)⁺] RNA, primed by a dT(16) oligomer. This mixture was heated to 70°C for 10 min, and then transferred to ice. A premixed solution, consisting of 200 U Superscript II (Gibco), buffer, deoxyribonucleoside triphosphates, and fluorescent nucleotides, was added to the RNA. Nucleotides were used at these final concentrations: 500 µM for dATP, dCTP, and dGTP and 200 µM for dTTP. Cy3-dUTP and Cy5-dUTP were used at a final concentration of 100 µM. The reaction was then incubated at 42°C for 2 hours. Unincorporated fluorescent nucleotides were removed by first diluting the reaction mixture with 470 µl of 10 mM tris-HCl (pH 8.0)/1 mM EDTA and then subsequently concentrating the mix to ~5 µl, using Centricon-30 microconcentrators (Amicon).
 12. Purified, labeled cDNA was resuspended in 11 µl of 3.5× SSC containing 10 µg poly(dA) and 0.3 µl of 10% SDS. Before hybridization, the solution was boiled for 2 min and then allowed to cool to room temperature. The solution was applied to the microarray under a cover slip, and the slide was placed in a custom hybridization chamber which was subsequently incubated for ~8 to 12 hours in a water bath at 62°C. Before scanning, slides were washed in 2× SSC, 0.2% SDS for 5 min, and then 0.05× SSC for 1 min. Slides were dried before scanning by centrifugation at 500 rpm in a Beckman CS-6R centrifuge.
 13. The complete data set is available on the Internet at cmgm.stanford.edu/pbrown/explore/index.html
 14. For 95% of all the genes analyzed, the mRNA levels measured in cells harvested at the first and second interval after inoculation differed by a factor of less than 1.5. The correlation coefficient for the comparison between mRNA levels measured for each gene in these two different mRNA samples was 0.98. When duplicate mRNA preparations from the same cell sample were compared in the same way, the correlation coefficient between the expression levels measured for the two samples by comparative hybridization was 0.99.
 15. The numbers and identities of known and putative genes, and their homologies to other genes, were gathered from the following public databases: *Saccharomyces* Genome Database (genome-www.stanford.edu), Yeast Protein Database (quest7.proteome.com), and Munich Information Centre for Protein Sequences (speedy.mips.biochem.mpg.de/mips/yeast/index.htm).
 16. A. Scholer and H. J. Schuller, *Mol. Cell. Biol.* 14, 3613 (1994).
 17. S. Kratzer and H. J. Schuller, *Gene* 161, 75 (1995).
 18. R. J. Haselbeck and H. L. McAlister, *J. Biol. Chem.* 268, 12116 (1993).
 19. M. Fernandez, E. Fernandez, R. Rodicio, *Mol. Gen. Genet.* 242, 727 (1994).
 20. A. Hartig et al., *Nucleic Acids Res.* 20, 5677 (1992).
 21. P. M. Martinez et al., *EMBO J.* 15, 2227 (1996).
 22. J. C. Varela, U. M. Praekelt, P. A. Meacock, R. J. Planta, W. H. Mager, *Mol. Cell. Biol.* 15, 6232 (1995).
 23. H. Ruis and C. Schuller, *Bioessays* 17, 959 (1995).
 24. J. L. Parrou, M. A. Teste, J. Francois, *Microbiology* 143, 1891 (1997).
 25. This expression profile was defined as having an induction of greater than 10-fold at 18.5 hours and less than 11-fold at 20.5 hours.
 26. S. L. Forsburg and L. Guarente, *Genes Dev.* 3, 1166 (1989).
 27. J. T. Olesen and L. Guarente, *ibid.* 4, 1714 (1990).
 28. M. Rosenkrantz, C. S. Kell, E. A. Pennell, L. J. Devenish, *Mol. Microbiol.* 13, 119 (1994).
 29. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr. The nucleotide codes are as follows: B-C, G, or T; N-G, A, T, or C; R-A or G; and Y-C or T.
 30. C. Fondrat and A. Kalogeropoulos, *Comput. Appl. Biosci.* 12, 363 (1996).
 31. D. Shore, *Trends Genet.* 10, 408 (1994).
 32. R. J. Planta and H. A. Raue, *ibid.* 4, 64 (1988).
 33. The degenerate consensus sequence VYCYRNNC-MNH was used to search for potential RAP1-binding sites. The exact consensus, as defined by (30), is WACAYCORTACATYW, with up to three differences allowed.
 34. S. F. Neuman, S. Bhattacharya, J. R. Broach, *Mol. Cell. Biol.* 15, 3187 (1995).
 35. P. Lesage, X. Yang, M. Carlson, *ibid.* 16, 1921 (1996).
 36. For example, we observed large inductions of the genes coding for *PCK1*, *FBP1* [Z. Yin et al., *Mol. Microbiol.* 20, 751 (1996)], the central glyoxylate cycle gene *ICL1* [A. Scholer and H. J. Schuller, *Curr. Genet.* 23, 375 (1993)], and the "aerobic" isoform of acetyl-CoA synthase, *ACS1* [M. A. van den Berg et al., *J. Biol. Chem.* 271, 28953 (1996)], with concomitant down-regulation of the glycolytic-specific genes *PFK1* and *PFK2* [P. A. Moore et al., *Mol. Cell. Biol.* 11, 5330 (1991)]. Other genes not directly involved in carbon metabolism but known to be induced upon nutrient limitation include genes encoding cytosolic catalase *CTT1* [P. H. Bissinger et al., *ibid.* 9, 1309 (1989)] and several genes encoding small heat-shock proteins, such as *HSP12*, *HSP26*, and *HSP42* [I. Farkas et al., *J. Biol. Chem.* 266, 15602 (1991); U. M. Praekelt and P. A. Meacock, *Mol. Gen. Genet.* 223, 97 (1990); D. Wotton et al., *J. Biol. Chem.* 271, 2717 (1996)].
 37. The levels of induction we measured for genes that were expressed at very low levels in the uninduced state (notably, *FBP1* and *PCK1*) were generally lower than those previously reported. This discrepancy was likely due to the conservative background subtraction method we used, which generally resulted in overestimation of very low expression levels (46).
 38. Cross-hybridization of highly related sequences can also occasionally obscure changes in gene expression, an important concern where members of gene families are functionally specialized and differentially regulated. The major alcohol dehydrogenase genes, *ADH1* and *ADH2*, share 88% nucleotide identity. Reciprocal regulation of these genes is an important feature of the diauxic shift, but was not observed in this experiment, presumably because of cross-hybridization of the fluorescent cDNAs representing these two genes. Nevertheless, we were able to detect differential expression of closely related isoforms of other enzymes, such as *HXK1/HXK2* (77% identical) [P. Herrero et al., *Yeast* 11, 137 (1995)], *MLS1/DAL7* (73% identical) (20), and *PGM1/PGM2* (72% identical) [D. Oh, J. E. Hopper, *Mol. Cell. Biol.* 10, 1415 (1990)], in accord with previous studies. Use in the microarray of deliberately selected DNA sequences corresponding to the most divergent segments of homologous genes, in lieu of the complete gene sequences, should relieve this problem in many cases.
 39. F. E. Williams, U. Varanasi, R. J. Trumbly, *Mol. Cell. Biol.* 11, 3307 (1991).
 40. D. Tzamaras and K. Struhl, *Nature* 369, 758 (1994).
 41. Differences in mRNA levels between the *tup1Δ* and wild-type strain were measured in two independent experiments. The correlation coefficient between the complete sets of expression ratios measured in these duplicate experiments was 0.83. The concordance between the sets of genes that appeared to be induced was very high between the two experiments. When only the 355 genes that showed at least a twofold increase in mRNA in the *tup1Δ* strain in either of the duplicate experiments were compared, the correlation coefficient was 0.82.
 42. The *tup1Δ* mutation consists of an insertion of the *LEU2* coding sequence, including a stop codon, between the ATG of *TUP1* and an Eco RI site 124 base pairs before the stop codon of the *TUP1* gene.
 43. L. R. Kowalski, K. Kondo, M. Inouye, *Mol. Microbiol.* 15, 341 (1995).
 44. M. Viswanathan, G. Muthukumar, Y. S. Cong, J. Lenard, *Gene* 148, 149 (1994).
 45. D. Hirata, K. Yano, T. Miyakawa, *Mol. Gen. Genet.* 242, 250 (1994).
 46. A. Gutierrez, L. Caramelo, A. Prieto, M. J. Martinez, A. T. Martinez, *Appl. Environ. Microbiol.* 60, 1783 (1994).
 47. A. Muheim et al., *Eur. J. Biochem.* 195, 369 (1991).
 48. J. A. Wemmie, M. S. Szczypka, D. J. Thiele, W. S. Moye-Rowley, *J. Biol. Chem.* 269, 32592 (1994).
 49. Microarrays were scanned using a custom-built scanning laser microscope built by S. Smith with software written by N. Ziv. Details concerning scanner design and construction are available at cmgm.stanford.edu/pbrown. Images were scanned at a resolution of 20 µm per pixel. A separate scan, using the appropriate excitation line, was done for each of the two fluorophores used. During the scanning process, the ratio between the signals in the two channels was calculated for several array elements containing total genomic DNA. To normalize the two channels with respect to overall intensity, we then adjusted photomultiplier and laser power settings such that the signal ratio at these elements was as close to 1.0 as possible. The combined images were analyzed with custom-written software. A bounding box, fitted to the size of the DNA spots in each quadrant, was placed over each array element. The average fluorescent intensity was calculated by summing the intensities of each pixel present in a bounding box, and then dividing by the total number of pixels. Local area background was calculated for each array element by determining the average fluorescent intensity for the lower 20% of pixel intensities. Although this method tends to underestimate the background, causing an underestimation of extreme ratios, it produces a very consistent and noise-tolerant approximation. Although the analog-to-digital board used for data collection possesses a wide dynamic range (12 bits), several signals were saturated (greater than the maximum signal intensity allowed) at the chosen settings. Therefore, extreme ratios at bright elements are generally underestimated. A signal was deemed significant if the average intensity after background subtraction was at least 2.5-fold higher than the standard deviation in the background measurements for all elements on the array.
 50. In addition to the 17 genes shown in Table 1, three additional genes were induced by an average of more than threefold in the duplicate experiments, but in one of the two experiments, the induction was less than twofold (range 1.6- to 1.9-fold).
 51. We thank H. Bennett, P. Spellman, J. Ravetto, M. Eisen, R. Pillai, B. Dunn, T. Ferea, and other members of the Brown lab for their assistance and helpful advice. We also thank S. Friend, D. Botstein, S. Smith, J. Hudson, and D. Dolginow for advice, support, and encouragement; K. Struhl and S. Chatterjee for the *Tup1* deletion strain; L. Fernandes for helpful advice on *Yap1*; and S. Klapholz and the reviewers for many helpful comments on the manuscript. Supported by a grant from the National Human Genome Research Institute (NHGRI) (HG00450), and by the Howard Hughes Medical Institute (HHMI). J.D.R. was supported by the HHMI and the NHGRI. V.R. was supported in part by an Institutional Training Grant in Genome Science (T32 HG00044) from the NHGRI. P.O.B. is an associate investigator of the HHMI.

5 September 1997; accepted 22 September 1997

Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships

STEVEN E. BRENNER*†‡, CYRUS CHOTHIA*, AND TIM J. P. HUBBARD§

*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom; and §Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, United Kingdom

Communicated by David R. Davies, National Institute of Diabetes, Bethesda, MD, March 16, 1998 (received for review November 12, 1997)

ABSTRACT Pairwise sequence comparison methods have been assessed using proteins whose relationships are known reliably from their structures and functions, as described in the SCOP database [Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia C. (1995) *J. Mol. Biol.* 247, 536–540]. The evaluation tested the programs BLAST [Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* 215, 403–410], WU-BLAST2 [Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* 266, 460–480], FASTA [Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* 85, 2444–2448], and SSEARCH [Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* 147, 195–197] and their scoring schemes. The error rate of all algorithms is greatly reduced by using statistical scores to evaluate matches rather than percentage identity or raw scores. The E-value statistical scores of SSEARCH and FASTA are reliable: the number of false positives found in our tests agrees well with the scores reported. However, the P-values reported by BLAST and WU-BLAST2 exaggerate significance by orders of magnitude. SSEARCH, FASTA ktup = 1, and WU-BLAST2 perform best, and they are capable of detecting almost all relationships between proteins whose sequence identities are >30%. For more distantly related proteins, they do much less well; only one-half of the relationships between proteins with 20–30% identity are found. Because many homologs have low sequence similarity, most distant relationships cannot be detected by any pairwise comparison method; however, those which are identified may be used with confidence.

Sequence database searching plays a role in virtually every branch of molecular biology and is crucial for interpreting the sequences issuing forth from genome projects. Given the method's central role, it is surprising that overall and relative capabilities of different procedures are largely unknown. It is difficult to verify algorithms on sample data because this requires large data sets of proteins whose evolutionary relationships are known unambiguously and independently of the methods being evaluated. However, nearly all known homologs have been identified by sequence analysis (the method to be tested). Also, it is generally very difficult to know, in the absence of structural data, whether two proteins that lack clear sequence similarity are unrelated. This has meant that although previous evaluations have helped improve sequence comparison, they have suffered from insufficient, imperfectly characterized, or artificial test data. Assessment also has been problematic because high quality database sequence searching attempts to have both sensitivity (detection of homologs) and specificity (rejection of unrelated proteins); however, these complementary goals are linked such that increasing one causes the other to be reduced.

Sequence comparison methodologies have evolved rapidly, so no previously published tests have evaluated modern versions of programs commonly used. For example, parameters in BLAST (1) have changed, and WU-BLAST2 (2)—which produces gapped alignments—has become available. The latest version of FASTA (3) previously tested was 1.6, but the current release (version 3.0) provides fundamentally different results in the form of statistical scoring.

The previous reports also have left gaps in our knowledge. For example, there has been no published assessment of thresholds for scoring schemes more sophisticated than percentage identity. Thus, the widely discussed statistical scoring measures have never actually been evaluated on large databases of real proteins. Moreover, the different scoring schemes commonly in use have not been compared.

Beyond these issues, there is a more fundamental question: in an absolute sense, how well does pairwise sequence comparison work? That is, what fraction of homologous proteins can be detected using modern database searching methods?

In this work, we attempt to answer these questions and to overcome both of the fundamental difficulties that have hindered assessment of sequence comparison methodologies. First, we use the set of distant evolutionary relationships in the SCOP: Structural Classification of Proteins database (4), which is derived from structural and functional characteristics (5). The SCOP database provides a uniquely reliable set of homologs, which are known independently of sequence comparison. Second, we use an assessment method that jointly measures both sensitivity and specificity. This method allows straightforward comparison of different sequence searching procedures. Further, it can be used to aid interpretation of real database searches and thus provide optimal and reliable results.

Previous Assessments of Sequence Comparison. Several previous studies have examined the relative performance of different sequence comparison methods. The most encompassing analyses have been by Pearson (6, 7), who compared the three most commonly used programs. Of these, the Smith–Waterman algorithm (8) implemented in SSEARCH (3) is the oldest and slowest but the most rigorous. Modern heuristics have provided BLAST (1) the speed and convenience to make it the most popular program. Intermediate between these two is FASTA (3), which may be run in two modes offering either greater speed (ktup = 2) or greater effectiveness (ktup = 1). Pearson also considered different parameters for each of these programs.

To test the methods, Pearson selected two representative proteins from each of 67 protein superfamilies defined by the PIR database (9). Each was used as a query to search the database, and the matched proteins were marked as being homologous or unrelated according to their membership of PIR

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. §1734 solely to indicate this fact.

© 1998 by The National Academy of Sciences 0027-8424/98/956073-6\$2.00/0 PNAS is available online at <http://www.pnas.org>.

Abbreviation: EPQ, errors per query.

†Present address: Department of Structural Biology, Stanford University, Fairchild Building D-109, Stanford, CA 94305-5126

‡To whom reprints requests should be addressed. e-mail: brenner@hyper.stanford.edu.

superfamilies. Pearson found that modern matrices and "ln-scaling" of raw scores improve results considerably. He also reported that the rigorous Smith–Waterman algorithm worked slightly better than FASTA, which was in turn more effective than BLAST.

Very large scale analyses of matrices have been performed (10), and Henikoff and Henikoff (11) also evaluated the effectiveness of BLAST and FASTA. Their test with BLAST considered the ability to detect homologs above a predetermined score but had no penalty for methods which also reported large numbers of spurious matches. The Henikoffs searched the SWISS-PROT database (12) and used PROSITE (13) to define homologous families. Their results showed that the BLOSUM62 matrix (14) performed markedly better than the extrapolated PAM-series matrices (15), which previously had been popular.

A crucial aspect of any assessment is the data that are used to test the ability of the program to find homologs. But in Pearson's and the Henikoffs' evaluations of sequence comparison, the correct results were effectively unknown. This is because the superfamilies in PIR and PROSITE are principally created by using the same sequence comparison methods which are being evaluated. Interdependency of data and methods creates a "chicken and egg" problem, and means for example, that new methods would be penalized for correctly identifying homologs missed by older programs. For instance, immunoglobulin variable and constant domains are clearly homologous, but PIR places them in different superfamilies. The problem is widespread: each superfamily in PIR 48.00 with a structural homolog is itself homologous to an average of 1.6 other PIR superfamilies (16).

To surmount these sorts of difficulties, Sander and Schneider (17) used protein structures to evaluate sequence comparison. Rather than comparing different sequence comparison algorithms, their work focused on determining a length-dependent threshold of percentage identity, above which all proteins would be of similar structure. A result of this analysis was the HSSP equation; it states that proteins with 25% identity over 80 residues will have similar structures, whereas shorter alignments require higher identity. (Other studies also have used structures (18–20), but these focused on a small number of model proteins and were principally oriented toward evaluating alignment accuracy rather than homology detection.)

A general solution to the problem of scoring comes from statistical measures (i.e., E-values and P-values) based on the extreme value distribution (21). Extreme value scoring was implemented analytically in the BLAST program using the Karlin and Altschul statistics (22, 23) and empirical approaches have been recently added to FASTA and SSEARCH. In addition to being heralded as a reliable means of recognizing significantly similar proteins (24, 25), the mathematical tractability of statistical scores "is a crucial feature of the BLAST algorithm" (1). The validity of this scoring procedure has been tested analytically and empirically (see ref. 2 and references in ref. 24). However, all large empirical tests used random sequences that may lack the subtle structure found within biological sequences (26, 27) and obviously do not contain any real homologs. Thus, although many researchers have suggested that statistical scores be used to rank matches (24, 25, 28), there have been no large rigorous experiments on biological data to determine the degree to which such rankings are superior.

A Database for Testing Homology Detection. Since the discovery that the structures of hemoglobin and myoglobin are very similar though their sequences are not (29), it has been apparent that comparing structures is a more powerful (if less convenient) way to recognize distant evolutionary relationships than comparing sequences. If two proteins show a high degree of similarity in their structural details and function, it

is very probable that they have an evolutionary relationship though their sequence similarity may be low.

The recent growth of protein structure information combined with the comprehensive evolutionary classification in the SCOP database (4, 5) have allowed us to overcome previous limitations. With these data, we can evaluate the performance of sequence comparison methods on real protein sequences whose relationships are known confidently. The SCOP database uses structural information to recognize distant homologs, the large majority of which can be determined unambiguously. These superfamilies, such as the globins or the immunoglobulins, would be recognized as related by the vast majority of the biological community despite the lack of high sequence similarity.

From SCOP, we extracted the sequences of domains of proteins in the Protein Data Bank (PDB) (30) and created two databases. One (PDB90D-B) has domains, which were all <90% identical to any other, whereas (PDB40D-B) had those <40% identical. The databases were created by first sorting all protein domains in SCOP by their quality and making a list. The highest quality domain was selected for inclusion in the database and removed from the list. Also removed from the list (and discarded) were all other domains above the threshold level of identity to the selected domain. This process was repeated until the list was empty. The PDB40D-B database contains 1,323 domains, which have 9,044 ordered pairs of distant relationships, or $\approx 0.5\%$ of the total 1,749,006 ordered pairs. In PDB90D-B, the 2,079 domains have 53,988 relationships, representing 1.2% of all pairs. Low complexity regions of sequence can achieve spurious high scores, so these were masked in both databases by processing with the SEG program (27) using recommended parameters: 12 1.8 2.0. The databases used in this paper are available from <http://sss.stanford.edu/sss/>, and databases derived from the current version of SCOP may be found at <http://scop.mrc-lmb.cam.ac.uk/scop/>.

Analyses from both databases were generally consistent, but PDB40D-B focuses on distantly related proteins and reduces the heavy overrepresentation in the PDB of a small number of families (31, 32), whereas PDB90D-B (with more sequences) improves evaluations of statistics. Except where noted otherwise, the distant homolog results here are from PDB40D-B. Although the precise numbers reported here are specific to the structural domain databases used, we expect the trends to be general.

Assessment Data and Procedure. Our assessment of sequence comparison may be divided into four different major categories of tests. First, using just a single sequence comparison algorithm at a time, we evaluated the effectiveness of different scoring schemes. Second, we assessed the reliability of scoring procedures, including an evaluation of the validity of statistical scoring. Third, we compared sequence comparison algorithms (using the optimal scoring scheme) to determine their relative performance. Fourth, we examined the distribution of homologs and considered the power of pairwise sequence comparison to recognize them. All of the analyses used the databases of structurally identified homologs and a new assessment criterion.

The analyses tested BLAST (1), version 1.4.9MP, and WU-BLAST2 (2), version 2.0a13MP. Also assessed was the FASTA package, version 3.0t76 (3), which provided FASTA and the SSEARCH implementation of Smith–Waterman (8). For SSEARCH and FASTA, we used BLOSUM45 with gap penalties $-12/-1$ (7, 16). The default parameters and matrix (BLOSUM62) were used for BLAST and WU-BLAST2.

The "Coverage Vs. Error" Plot. To test a particular protocol (comprising a program and scoring scheme), each sequence from the database was used as a query to search the database. This yielded ordered pairs of query and target sequences with associated scores, which were sorted, on the basis of their scores, from best to worst. The ideal method would have

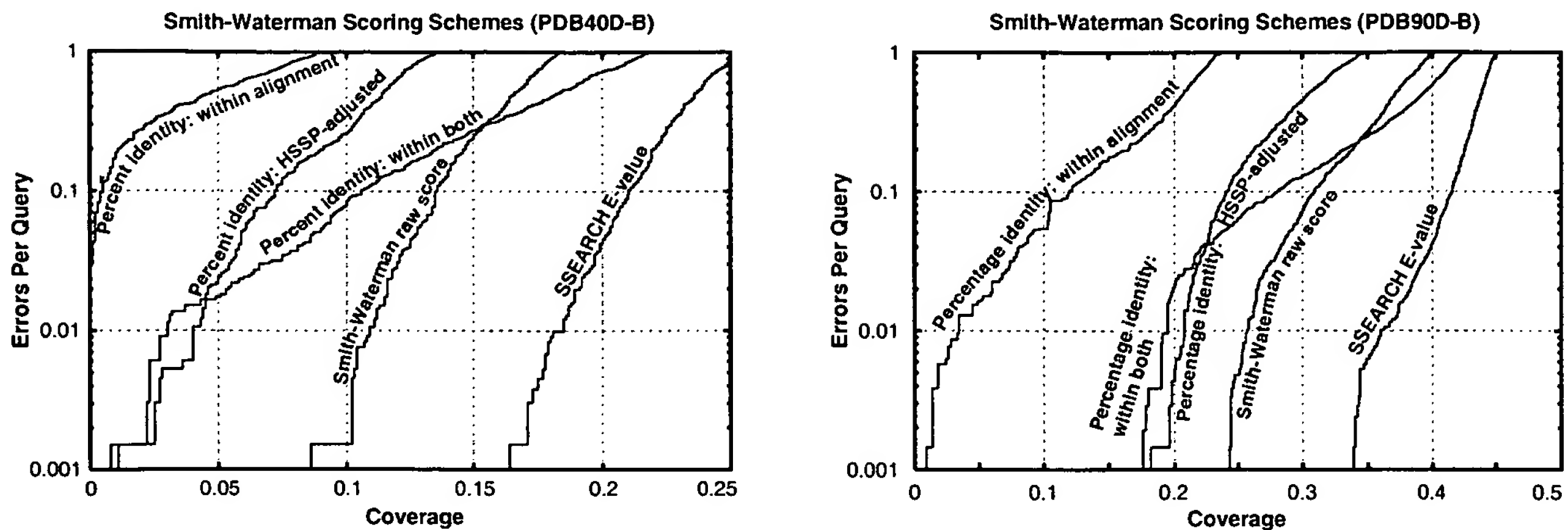


FIG. 1. Coverage vs. error plots of different scoring schemes for SSEARCH Smith-Waterman. (A) Analysis of PDB40D-B database. (B) Analysis of PDB90D-B database. All of the proteins in the database were compared with each other using the SSEARCH program. The results of this single set of comparisons were considered using five different scoring schemes and assessed. The graphs show the coverage and errors per query (EPQ) for statistical scores, raw scores, and three measures using percentage identity. In the coverage vs. error plot, the x axis indicates the fraction of all homologs in the database (known from structure) which have been detected. Precisely, it is the number of detected pairs of proteins with the same fold divided by the total number of pairs from a common superfamily. PDB40D-B contains a total of 9,044 homologs, so a score of 10% indicates identification of 904 relationships. The y axis reports the number of EPQ. Because there are 1,323 queries made in the PDB40D-B all-vs.-all comparison, 13 errors corresponds to 0.01, or 1% EPQ. The y axis is presented on a log scale to show results over the widely varying degrees of accuracy which may be desired. The scores that correspond to the levels of EPQ and coverage are shown in Fig. 4 and Table 1. The graph demonstrates the trade-off between sensitivity and selectivity. As more homologs are found (moving to the right), more errors are made (moving up). The ideal method would be in the lower right corner of the graph, which corresponds to identifying many evolutionary relationships without selecting unrelated proteins. Three measures of percentage identity are plotted. Percentage identity within alignment is the degree of identity within the aligned region of the proteins, without consideration of the alignment length. Percentage identity within both is the number of identical residues in the aligned region as a percentage of the average length of the query and target proteins. The HSP equation (17) is $H = 290.15l^{-0.562}$ where l is length for $10 < l < 80$; $H > 100$ for $l < 10$; $H = 24.7$ for $l > 80$. The percentage identity HSP-adjusted score is the percent identity within the alignment minus H . Smith-Waterman raw scores and E-values were taken directly from the sequence comparison program.

perfect separation, with all of the homologs at the top of the list and unrelated proteins below. In practice, perfect separation is impossible to achieve so instead one is interested in drawing a threshold above which there are the largest number of related pairs of sequences consistent with an acceptable error rate.

Our procedure involved measuring the coverage and error for every threshold. Coverage was defined as the fraction of structurally determined homologs that have scores above the selected threshold; this reflects the sensitivity of a method. Errors per query (EPQ), an indicator of selectivity, is the number of nonhomologous pairs above the threshold divided by the number of queries. Graphs of these data, called coverage vs. error plots, were devised to understand how

protocols compare at different levels of accuracy. These graphs share effectively all of the beneficial features of Receiver Operating Characteristic (ROC) plots (33, 34) but better represent the high degrees of accuracy required in sequence comparison and the huge background of nonhomologs.

This assessment procedure is directly relevant to practical sequence database searching, for it provides precisely the information necessary to perform a reliable sequence database search. The EPQ measure places a premium on score consistency; that is, it requires scores to be comparable for different queries. Consistency is an aspect which has been largely

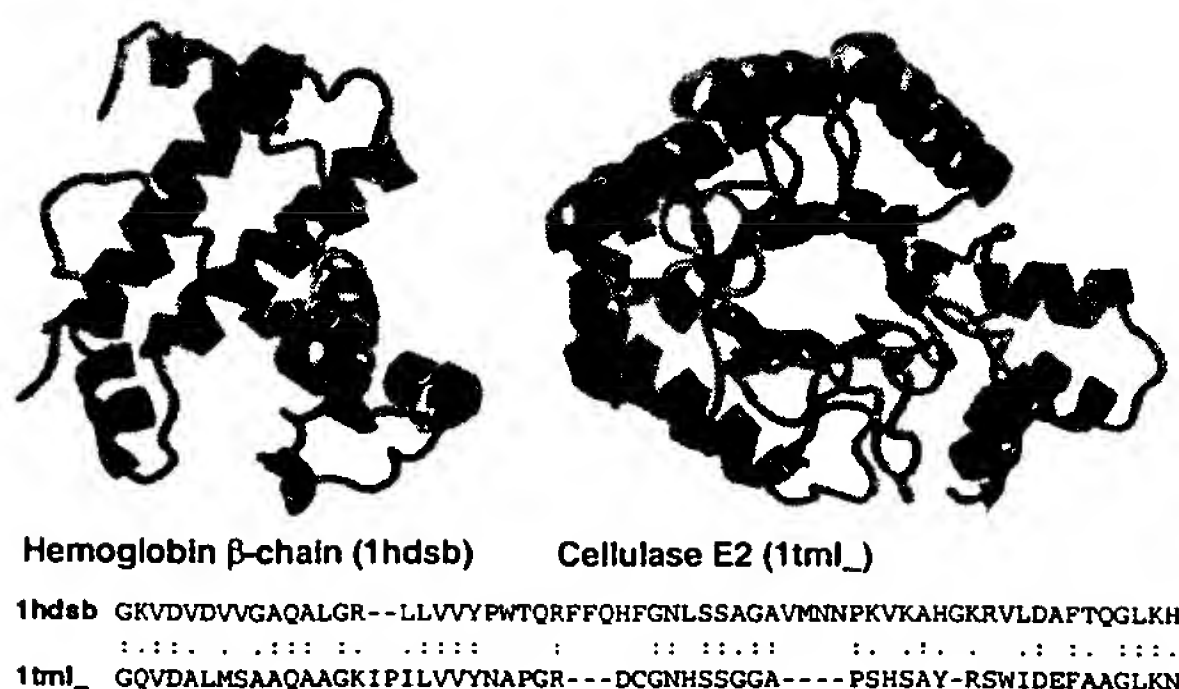


FIG. 2. Unrelated proteins with high percentage identity. Hemoglobin β -chain (PDB code 1hds chain b, ref. 38, *Left*) and cellulase E2 (PDB code 1tml, ref. 39, *Right*) have 39% identity over 64 residues, a level which is often believed to be indicative of homology. Despite this high degree of identity, their structures strongly suggest that these proteins are not related. Appropriately, neither the raw alignment score of 85 nor the E-value of 1.3 is significant. Proteins rendered by RASMOL (40).

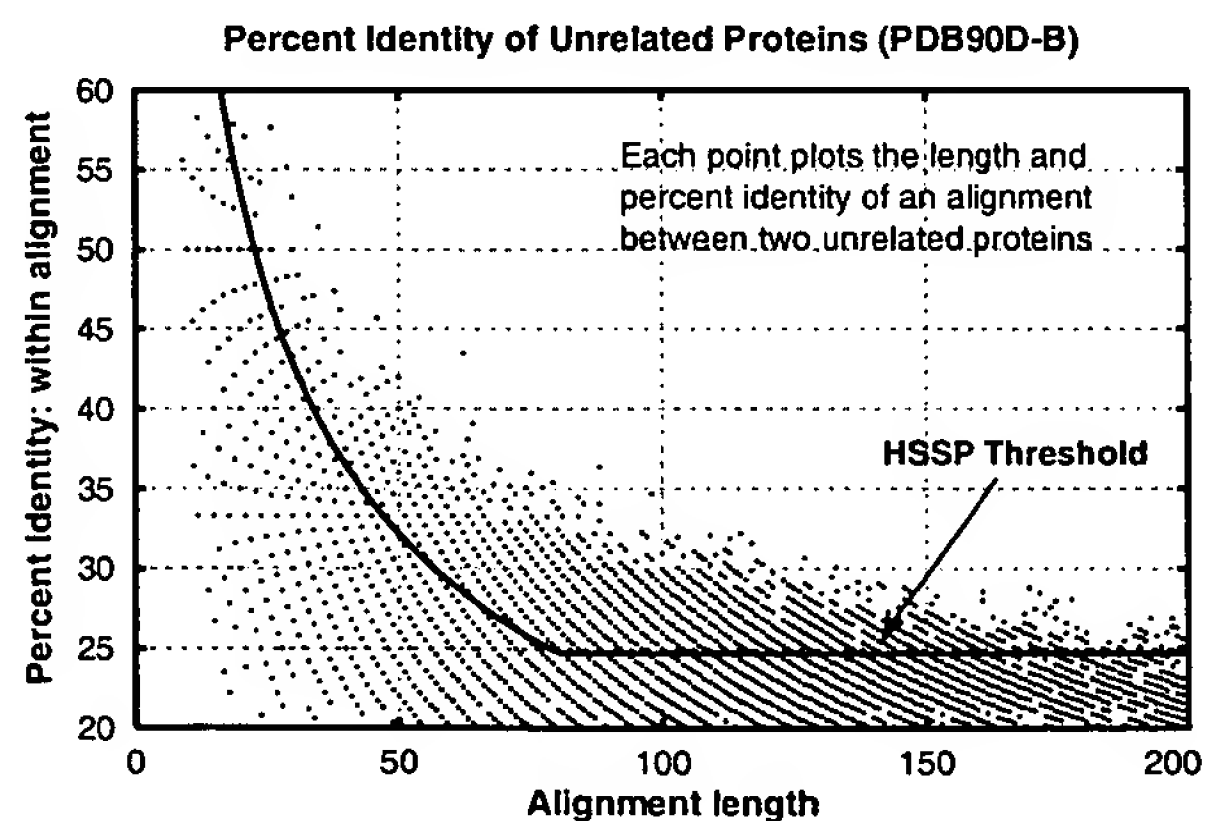


FIG. 3. Length and percentage identity of alignments of unrelated proteins in PDB90D-B: Each pair of nonhomologous proteins found with SSEARCH is plotted as a point whose position indicates the length and the percentage identity within the alignment. Because alignment length and percentage identity are quantized, many pairs of proteins may have exactly the same alignment length and percentage identity. The line shows the HSP threshold (though it is intended to be applied with a different matrix and parameters).

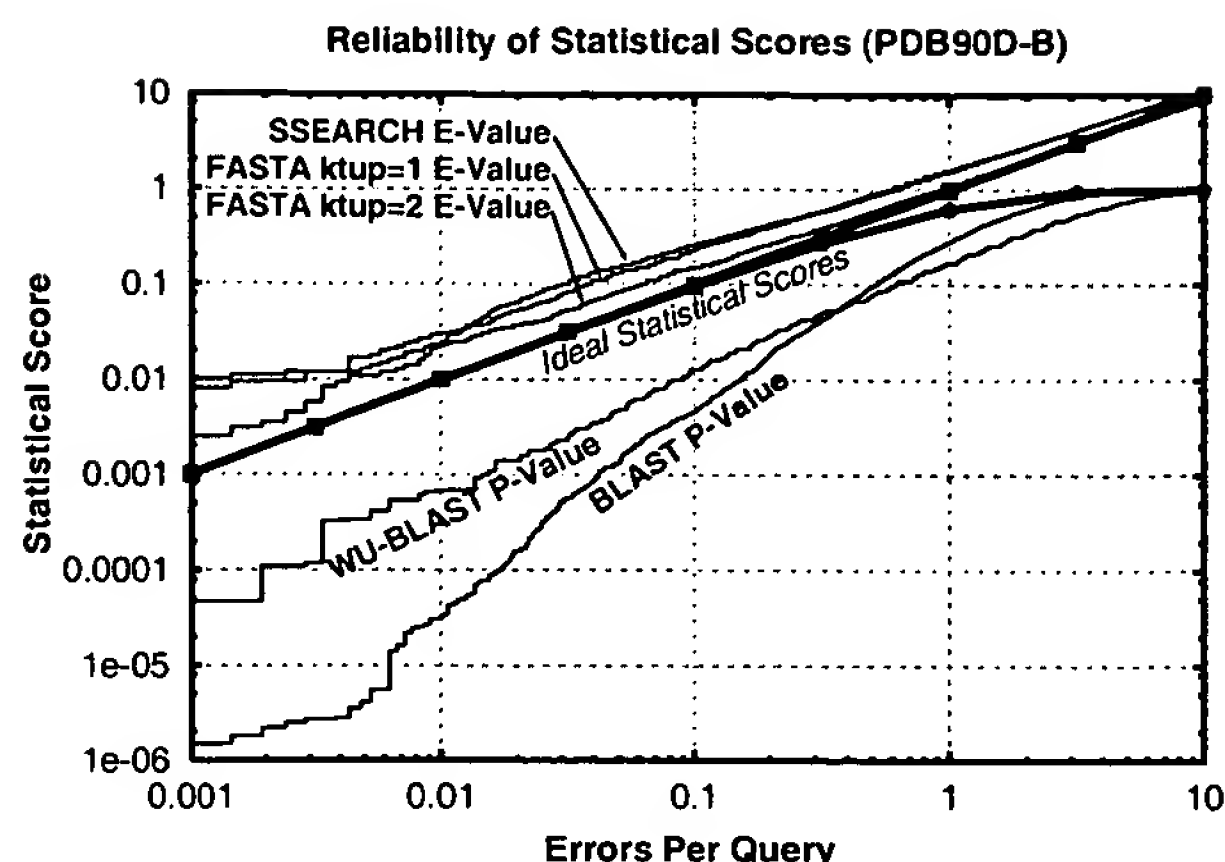


FIG. 4. Reliability of statistical scores in PDB90D-B: Each line shows the relationship between reported statistical score and actual error rate for a different program. E-values are reported for SSEARCH and FASTA, whereas P-values are shown for BLAST and WU-BLAST2. If the scoring were perfect, then the number of errors per query and the E-values would be the same, as indicated by the upper bold line. (P-values should be the same as EPQ for small numbers, and diverges at higher values, as indicated by the lower bold line.) E-values from SSEARCH and FASTA are shown to have good agreement with EPQ but underestimate the significance slightly. BLAST and WU-BLAST2 are overconfident, with the degree of exaggeration dependent upon the score. The results for PDB40D-B were similar to those for PDB90D-B despite the difference in number of homologs detected. This graph could be used to roughly calibrate the reliability of a given statistical score.

ignored in previous tests but is essential for the straightforward or automatic interpretation of sequence comparison results. Further, it provides a clear indication of the confidence that should be ascribed to each match. Indeed, the EPQ measure should approximate the expectation value reported by database searching programs, if the programs' estimates are accurate.

The Performance of Scoring Schemes. All of the programs tested could provide three fundamental types of scores. The first score is the percentage identity, which may be computed in several ways based on either the length of the alignment or the lengths of the sequences. The second is a "raw" or "Smith-Waterman" score, which is the measure optimized by the Smith-Waterman algorithm and is computed by summing the substitution matrix scores for each position in the alignment and subtracting gap penalties. In BLAST, a measure

related to this score is scaled into bits. Third is a statistical score based on the extreme value distribution. These results are summarized in Fig. 1.

Sequence Identity. Though it has been long established that percentage identity is a poor measure (35), there is a common rule-of-thumb stating that 30% identity signifies homology. Moreover, publications have indicated that 25% identity can be used as a threshold (17, 36). We find that these thresholds, originally derived years ago, are not supported by present results. As databases have grown, so have the possibilities for chance alignments with high identity; thus, the reported cutoffs lead to frequent errors. Fig. 2 shows one of the many pairs of proteins with very different structures that nonetheless have high levels of identity over considerable aligned regions. Despite the high identity, the raw and the statistical scores for such incorrect matches are typically not significant. The principal reasons percentage identity does so poorly seem to be that it ignores information about gaps and about the conservative or radical nature of residue substitutions.

From the PDB90D-B analysis in Fig. 3, we learn that 30% identity is a reliable threshold for this database only for sequence alignments of at least 150 residues. Because one unrelated pair of proteins has 43.5% identity over 62 residues, it is probably necessary for alignments to be at least 70 residues in length before 40% is a reasonable threshold, for a database of this particular size and composition.

At a given reliability, scores based on percentage identity detect just a fraction of the distant homologs found by statistical scoring. If one measures the percentage identity in the aligned regions without consideration of alignment length, then a negligible number of distant homologs are detected. Use of the HSP equation improves the value of percentage identity, but even this measure can find only 4% of all known homologs at 1% EPQ. In short, percentage identity discards most of the information measured in a sequence comparison.

Raw Scores. Smith-Waterman raw scores perform better than percentage identity (Fig. 1), but ln-scaling (7) provided no notable benefit in our analysis. It is necessary to be very precise when using either raw or bit scores because a 20% change in cutoff score could yield a tenfold difference in EPQ. However, it is difficult to choose appropriate thresholds because the reliability of a bit score depends on the lengths of the proteins matched and the size of the database. Raw score thresholds also are affected by matrix and gap parameters.

Statistical Scores. Statistical scores were introduced partly to overcome the problems that arise from raw scores. This scoring scheme provides the best discrimination between homologous proteins and those which are unrelated. Most

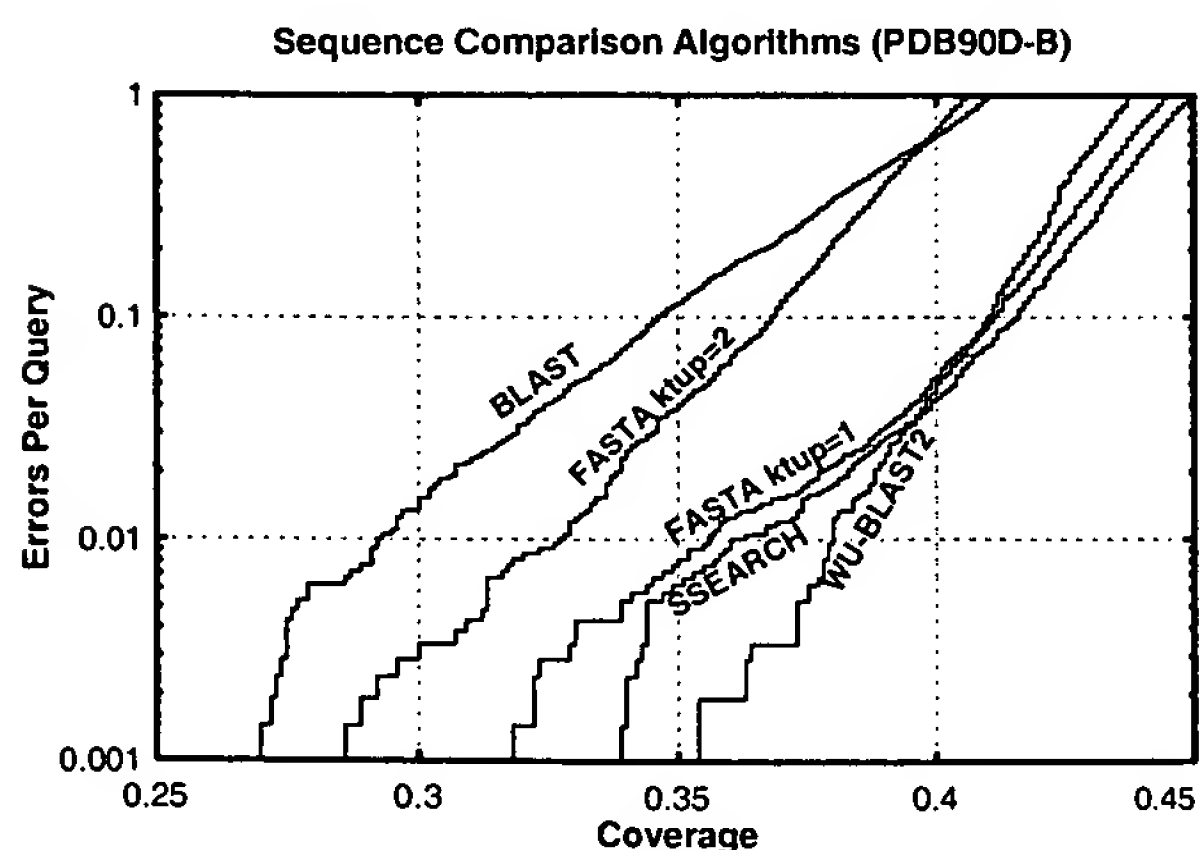
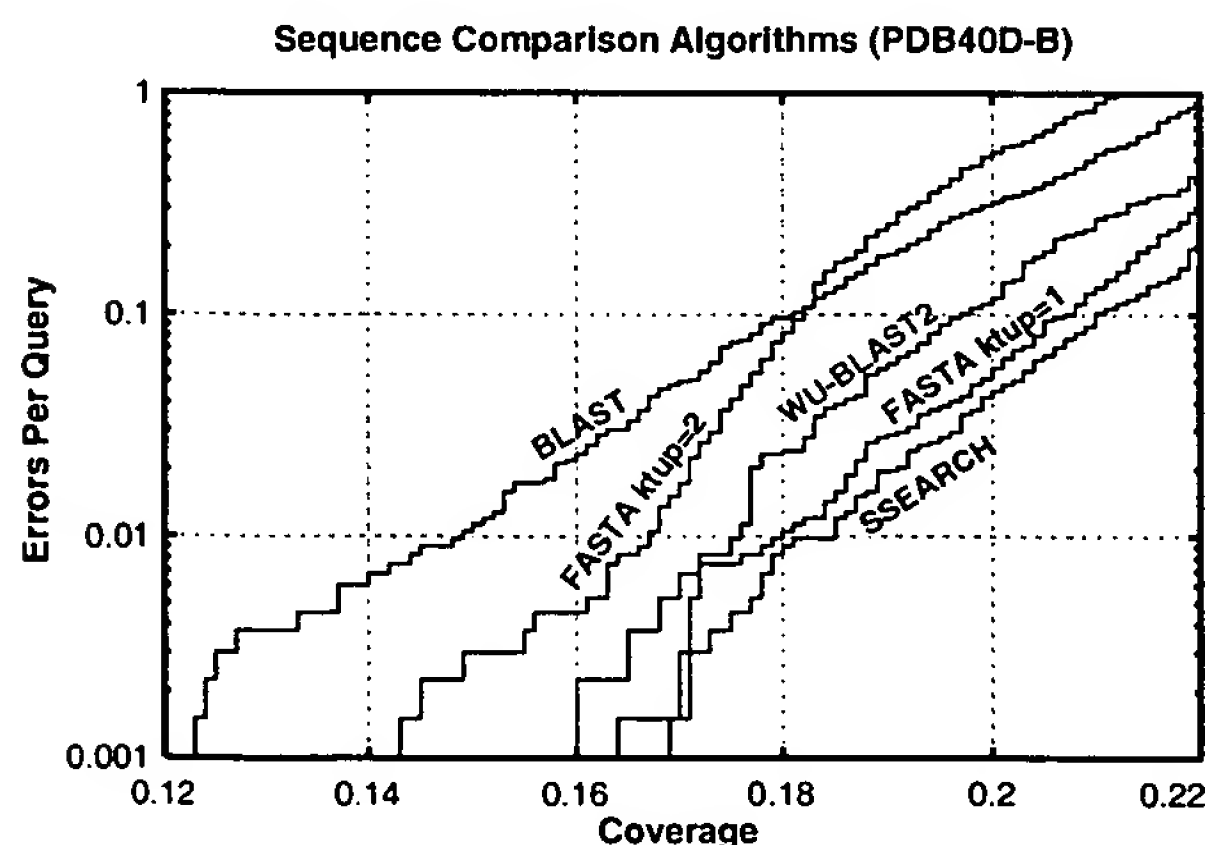


FIG. 5. Coverage vs. error plots of different sequence comparison methods: Five different sequence comparison methods are evaluated, each using statistical scores (E- or P-values). (A) PDB40D-B database. In this analysis, the best method is the slow SSEARCH, which finds 18% of relationships at 1% EPQ. FASTA ktup = 1 and WU-BLAST2 are almost as good. (B) PDB90D-B database. The quick WU-BLAST2 program provides the best coverage at 1% EPQ on this database, although at higher levels of error it becomes slightly worse than FASTA ktup = 1 and SSEARCH.

likely, its power can be attributed to its incorporation of more information than any other measure; it takes account of the full substitution and gap data (like raw scores) but also has details about the sequence lengths and composition and is scaled appropriately.

We find that statistical scores are not only powerful, but also easy to interpret. SSEARCH and FASTA show close agreement between statistical scores and actual number of errors per query (Fig. 4). The expectation value score gives a good, slightly conservative estimate of the chances of the two sequences being found at random in a given query. Thus, an E-value of 0.01 indicates that roughly one pair of nonhomologs of this similarity should be found in every 100 different queries. Neither raw scores nor percentage identity can be interpreted in this way, and these results validate the suitability of the extreme value distribution for describing the scores from a database search.

The P-values from BLAST also should be directly interpretable but were found to overstate significance by more than two orders of magnitude for 1% EPQ for this database. Nonetheless, these results strongly suggest that the analytic theory is fundamentally appropriate. WU-BLAST2 scores were more reliable than those from BLAST, but also exaggerate expected confidence by more than an order of magnitude at 1% EPQ.

Overall Detection of Homologs and Comparison of Algorithms. The results in Fig. 5A and Table 1 show that pairwise sequence comparison is capable of identifying only a small fraction of the homologous pairs of sequences in PDB40D-B. Even SSEARCH with E-values, the best protocol tested, could find only 18% of all relationships at a 1% EPQ. BLAST, which identifies 15%, was the worst performer, whereas FASTA $k_{\text{tup}} = 1$ is nearly as effective as SSEARCH. FASTA $k_{\text{tup}} = 2$ and WU-BLAST2 are intermediate in their ability to detect homologs. Comparison of different algorithms indicates that those capable of identifying more homologs are generally slower. SSEARCH is 25 times slower than BLAST and 6.5 times slower than FASTA $k_{\text{tup}} = 1$. WU-BLAST2 is slightly faster than FASTA $k_{\text{tup}} = 2$, but the latter has more interpretable scores.

In PDB90D-B, where there are many close relationships, the best method can identify only 38% of structurally known homologs (Fig. 5B). The method which finds that many relationships is WU-BLAST2. Consequently, we infer that the differences between FASTA $k_{\text{tup}} = 1$, SSEARCH, and WU-BLAST2 programs are unlikely to be significant when compared with variation in database composition and scoring reliability.

Fig. 6 helps to explain why most distant homologs cannot be found by sequence comparison: a great many such relationships have no more sequence identity than would be expected by chance. SSEARCH with E-values can recognize >90% of the homologous pairs with 30–40% identity. In this region, there are 30 pairs of homologous proteins that do not have significant E-values, but 26 of these involve sequences with <50 residues. Of sequences having 25–30% identity, 75% are identified by SSEARCH E-values. However, although the number of homologs grows at lower levels of identity, the detection falls off sharply: only 40% of homologs with 20–25% identity

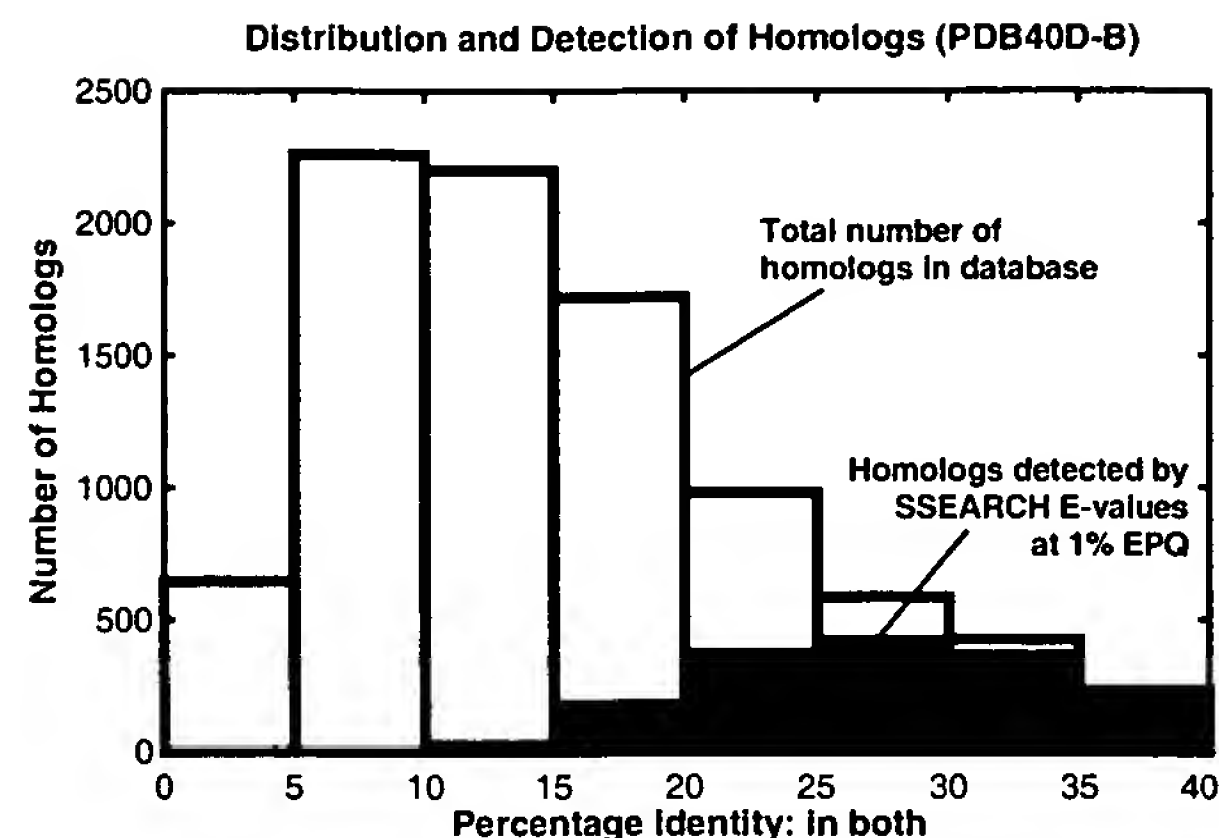


FIG. 6. Distribution and detection of homologs in PDB40D-B. Bars show the distribution of homologous pairs PDB40D-B according to their identity (using the measure of identity in both). Filled regions indicate the number of these pairs found by the best database searching method (SSEARCH with E-values) at 1% EPQ. The PDB40D-B database contains proteins with <40% identity, and as shown on this graph, most structurally identified homologs in the database have diverged extremely far in sequence and have <20% identity. Note that the alignments may be inaccurate, especially at low levels of identity. Filled regions show that SSEARCH can identify most relationships that have 25% or more identity, but its detection wanes sharply below 25%. Consequently, the great sequence divergence of most structurally identified evolutionary relationships effectively defeats the ability of pairwise sequence comparison to detect them.

are detected and only 10% of those with 15–20% can be found. These results show that statistical scores can find related proteins whose identity is remarkably low; however, the power of the method is restricted by the great divergence of many protein sequences.

After completion of this work, a new version of pairwise BLAST was released: BLASTGP (37). It supports gapped alignments, like WU-BLAST2, and dispenses with sum statistics. Our initial tests on BLASTGP using default parameters show that its E-values are reliable and that its overall detection of homologs was substantially better than that of ungapped BLAST, but not quite equal to that of WU-BLAST2.

CONCLUSION

The general consensus amongst experts (see refs. 7, 24, 25, 27 and references therein) suggests that the most effective sequence searches are made by (i) using a large current database in which the protein sequences have been complexity masked and (ii) using statistical scores to interpret the results. Our experiments fully support this view.

Our results also suggest two further points. First, the E-values reported by FASTA and SSEARCH give fairly accurate estimates of the significance of each match, but the P-values provided by BLAST and WU-BLAST2 underestimate the true

Table 1. Summary of sequence comparison methods with PDB40D-B

Method	Relative Time*	1% EPQ Cutoff	Coverage at 1% EPQ
SSEARCH % identity: within alignment	25.5	>70%	<0.1
SSEARCH % identity: within both	25.5	34%	3.0
SSEARCH % identity: HSSP-scaled	25.5	35% (HSSP + 9.8)	4.0
SSEARCH Smith–Waterman raw scores	25.5	142	10.5
SSEARCH E-values	25.5	0.03	18.4
FASTA $k_{\text{tup}} = 1$ E-values	3.9	0.03	17.9
FASTA $k_{\text{tup}} = 2$ E-values	1.4	0.03	16.7
WU-BLAST2 P-values	1.1	0.003	17.5
BLAST P-values	1.0	0.00016	14.8

*Times are from large database searches with genome proteins.

extent of errors. Second, SSEARCH, WU-BLAST2, and FASTA ktup = 1 perform best, though BLAST and FASTA ktup = 2 detect most of the relationships found by the best procedures and are appropriate for rapid initial searches.

The homologous proteins that are found by sequence comparison can be distinguished with high reliability from the huge number of unrelated pairs. However, even the best database searching procedures tested fail to find the large majority of distant evolutionary relationships at an acceptable error rate. Thus, if the procedures assessed here fail to find a reliable match, it does not imply that the sequence is unique; rather, it indicates that any relatives it might have are distant ones.**

**Additional and updated information about this work, including supplementary figures, may be found at <http://sss.stanford.edu/sss/>.

The authors are grateful to Drs. A. G. Murzin, M. Levitt, S. R. Eddy, and G. Mitchison for valuable discussion. S.E.B. was principally supported by a St. John's College (Cambridge, UK) Benefactors' Scholarship and by the American Friends of Cambridge University. S.E.B. dedicates his contribution to the memory of Rabbi Albert T. and Clara S. Bilgray.

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215**, 403–410.
2. Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* **266**, 460–480.
3. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448.
4. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* **247**, 536–540.
5. Brenner, S. E., Chothia, C., Hubbard, T. J. P. & Murzin, A. G. (1996) *Methods Enzymol.* **266**, 635–643.
6. Pearson, W. R. (1991) *Genomics* **11**, 635–650.
7. Pearson, W. R. (1995) *Protein Sci.* **4**, 1145–1160.
8. Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **147**, 195–197.
9. George, D. G., Hunt, L. T. & Barker, W. C. (1996) *Methods Enzymol.* **266**, 41–59.
10. Vogt, G., Etzold, T. & Argos, P. (1995) *J. Mol. Biol.* **249**, 816–831.
11. Henikoff, S. & Henikoff, J. G. (1993) *Proteins* **17**, 49–61.
12. Bairoch, A. & Apweiler, R. (1996) *Nucleic Acids Res.* **24**, 21–25.
13. Bairoch, A., Bucher, P. & Hofmann, K. (1996) *Nucleic Acids Res.* **24**, 189–196.
14. Henikoff, S. & Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 10915–10919.
15. Dayhoff, M., Schwartz, R. M. & Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. (National Bio-medical Research Foundation, Silver Spring, MD), Vol. 5, Suppl. 3, pp. 345–352.
16. Brenner, S. E. (1996) Ph.D. thesis. (University of Cambridge, UK).
17. Sander, C. & Schneider, R. (1991) *Proteins* **9**, 56–68.
18. Johnson, M. S. & Overington, J. P. (1993) *J. Mol. Biol.* **233**, 716–738.
19. Barton, G. J. & Sternberg, M. J. E. (1987) *Protein Eng.* **1**, 89–94.
20. Lesk, A. M., Levitt, M. & Chothia, C. (1986) *Protein Eng.* **1**, 77–78.
21. Arratia, R., Gordon, L. & M, W. (1986) *Ann. Stat.* **14**, 971–993.
22. Karlin, S. & Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 2264–2268.
23. Karlin, S. & Altschul, S. F. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 5873–5877.
24. Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994) *Nat. Genet.* **6**, 119–129.
25. Pearson, W. R. (1996) *Methods Enzymol.* **266**, 227–258.
26. Lipman, D. J., Wilbur, W. J., Smith, T. F. & Waterman, M. S. (1984) *Nucleic Acids Res.* **12**, 215–226.
27. Wootton, J. C. & Federhen, S. (1996) *Methods Enzymol.* **266**, 554–571.
28. Waterman, M. S. & Vingron, M. (1994) *Stat. Science* **9**, 367–381.
29. Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965) *J. Mol. Biol.* **13**, 669–678.
30. Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987) in *Crystallographic Databases: Information Content, Software Systems, Scientific Applications*, eds. Allen, F. H., Bergerhoff, G. & Sievers, R. (Data Comm. Intl. Union Crystallogr., Cambridge, UK), pp. 107–132.
31. Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1997) *Curr. Opin. Struct. Biol.* **7**, 369–376.
32. Orengo, C., Michie, A., Jones S, Jones D. T, Swindells M. B. & Thornton, J. (1997) *Structure (London)* **5**, 1093–1108.
33. Zweig, M. H. & Campbell, G. (1993) *Clin. Chem.* **39**, 561–577.
34. Gribskov, M. & Robinson, N. L. (1996) *Comput. Chem.* **20**, 25–33.
35. Fitch, W. M. (1966) *J. Mol. Biol.* **16**, 9–16.
36. Chung, S. Y. & Subbiah, S. (1996) *Structure (London)* **4**, 1123–1127.
37. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25**, 3389–3402.
38. Girling, R., Schmidt, W., Jr, Houston, T., Amma, E. & Huisman, T. (1979) *J. Mol. Biol.* **131**, 417–433.
39. Spezio, M., Wilson, D. & Karplus, P. (1993) *Biochemistry* **32**, 9906–9916.
40. Sayle, R. A. & Milner-White, E. J. (1995) *Trends Biochem. Sci.* **20**, 374–376.

Differential gene expression in drug metabolism and toxicology: practicalities, problems and potential

JOHN C. ROCKETT†, DAVID J. ESDAILE‡
and G. GORDON GIBSON*

Molecular Toxicology Laboratory, School of Biological Sciences, University of Surrey,
Guildford, Surrey, GU2 5XH, UK

Received January 8, 1999

1. An important feature of the work of many molecular biologists is identifying which genes are switched on and off in a cell under different environmental conditions or subsequent to xenobiotic challenge. Such information has many uses, including the deciphering of molecular pathways and facilitating the development of new experimental and diagnostic procedures. However, the student of gene hunting should be forgiven for perhaps becoming confused by the mountain of information available as there appears to be almost as many methods of discovering differentially expressed genes as there are research groups using the technique.

2. The aim of this review was to clarify the main methods of differential gene expression analysis and the mechanistic principles underlying them. Also included is a discussion on some of the practical aspects of using this technique. Emphasis is placed on the so-called 'open' systems, which require no prior knowledge of the genes contained within the study model. Whilst these will eventually be replaced by 'closed' systems in the study of human, mouse and other commonly studied laboratory animals, they will remain a powerful tool for those examining less fashionable models.

3. The use of suppression-PCR subtractive hybridization is exemplified in the identification of up- and down-regulated genes in rat liver following exposure to phenobarbital, a well-known inducer of the drug metabolizing enzymes.

4. Differential gene display provides a coherent platform for building libraries and microchip arrays of 'gene fingerprints' characteristic of known enzyme inducers and xenobiotic toxicants, which may be interrogated subsequently for the identification and characterization of xenobiotics of unknown biological properties.

Introduction

It is now apparent that the development of almost all cancers and many non-neoplastic diseases are accompanied by altered gene expression in the affected cells compared to their normal state (Hunter 1991, Wynford-Thomas 1991, Vogelstein and Kinzler 1993, Semenza 1994, Cassidy 1995, Kleinjan and Van Hegningen 1998). Such changes also occur in response to external stimuli such as pathogenic micro-organisms (Rohn *et al.* 1996, Singh *et al.* 1997, Griffin and Krishna 1998, Lunney 1998) and xenobiotics (Sewall *et al.* 1995, Dogra *et al.* 1998, Ramana and Kohli 1998), as well as during the development of undifferentiated cells (Hecht 1998, Rudin and Thompson 1998, Schneider-Maunoury *et al.* 1998). The potential medical and therapeutic benefits of understanding the molecular changes which occur in any given cell in progressing from the normal to the 'altered' state are enormous. Such profiling essentially provides a 'fingerprint' of each step of a

* Author for correspondence; e-mail: g.gibson@surrey.ac.uk

† Current Address: US Environmental Protection Agency, National Health and Environmental Effects, Research Laboratory, Reproductive Toxicology Division, Research Triangle Park, NC 27711, USA.

‡ Rhone-Poulenc Agrochemicals, Toxicology Department, Sophia-Antipolis, Nice, France.

cell's development or response and should help in the elucidation of specific and sensitive biomarkers representing, for example, different types of cancer or previous exposure to certain classes of chemicals that are enzyme inducers.

In drug metabolism, many of the xenobiotic-metabolizing enzymes (including the well-characterized isoforms of cytochrome P450) are inducible by drugs and chemicals in man (Pelkonen *et al.* 1998), predominantly involving transcriptional activation of not only the cognate cytochrome P450 genes, but additional cellular proteins which may be crucial to the phenomenon of induction. Accordingly, the development of methodology to identify and assess the full complement of genes that are either up- or down-regulated by inducers are crucial in the development of knowledge to understand the precise molecular mechanisms of enzyme induction and how this relates to drug action. Similarly, in the field of chemical-induced toxicity, it is now becoming increasingly obvious that most adverse reactions to drugs and chemicals are the result of multiple gene regulation, some of which are causal and some of which are casually-related to the toxicological phenomenon *per se*. This observation has led to an upsurge in interest in gene-profiling technologies which differentiate between the control and toxin-treated gene pools in target tissues and is, therefore, of value in rationalizing the molecular mechanisms of xenobiotic-induced toxicity. Knowledge of toxin-dependent gene regulation in target tissues is not solely an academic pursuit as much interest has been generated in the pharmaceutical industry to harness this technology in the early identification of toxic drug candidates, thereby shortening the developmental process and contributing substantially to the safety assessment of new drugs. For example, if the gene profile in response to say a testicular toxin that has been well-characterized *in vivo* could be determined in the testis, then this profile would be representative of all new drug candidates which act via this specific molecular mechanism of toxicity, thereby providing a useful and coherent approach to the early detection of such toxicants. Whereas it would be informative to know the identity and functionality of all genes up/down regulated by such toxicants, this would appear a longer term goal, as the majority of human genes have not yet been sequenced, far less their functionality determined. However, the current use of gene profiling yields a *pattern* of gene changes for a xenobiotic of unknown toxicity which may be matched to that of well-characterized toxins, thus alerting the toxicologist to possible *in vivo* similarities between the unknown and the standard, thereby providing a platform for more extensive toxicological examination. Such approaches are beginning to gain momentum, in that several biotechnology companies are commercially producing 'gene chips' or 'gene arrays' that may be interrogated for toxicity assessment of xenobiotics. These chips consist of hundreds/thousands of genes, some of which are degenerate in the sense that not all of the genes are mechanistically-related to any one toxicological phenomenon. Whereas these chips are useful in broad-spectrum screening, they are maturing at a substantial rate, in that gene arrays are now becoming more specific, e.g. chips for the identification of changes in growth factor families that contribute to the aetiology and development of chemically-induced neoplasias.

Although documenting and explaining these genetic changes presents a formidable obstacle to understanding the different mechanisms of development and disease progression, the technology is now available to begin attempting this difficult challenge. Indeed, several 'differential expression analysis' methods have been developed which facilitate the identification of gene products that demonstrate

altered expression in cells of one population compared to another. These methods have been used to identify differential gene expression in many situations, including invading pathogenic microbes (Zhao *et al.* 1998), in cells responding to extracellular and intracellular microbial invasion (Duguid and Dinauer 1990, Ragno *et al.* 1997, Maldarelli *et al.* 1998), in chemically treated cells (Syed *et al.* 1997, Rockett *et al.* 1999), neoplastic cells (Liang *et al.* 1992, Chang and Terzaghi-Howe 1998), activated cells (Gurskaya *et al.* 1996, Wan *et al.* 1996), differentiated cells (Hara *et al.* 1991, Guimaraes *et al.* 1995a, b), and different cell types (Davis *et al.* 1984, Hedrick *et al.* 1984, Xhu *et al.* 1998). Although differential expression analysis technologies are applicable to a broad range of models, perhaps their most important advantage is that, in most cases, absolutely no prior knowledge of the specific genes which are up- or down-regulated is required.

The field of differential expression analysis is a large and complex one, with many techniques available to the potential user. These can be categorized into several methodological approaches, including:

- (1) Differential screening,
- (2) Subtractive hybridization (SH) (includes methods such as chemical cross-linking subtraction—CCLS, suppression-PCR subtractive hybridization—SSH, and representational difference analysis—RDA),
- (3) Differential display (DD),
- (4) Restriction endonuclease facilitated analysis (including serial analysis of gene expression—SAGE—and gene expression fingerprinting—GEF),
- (5) Gene expression arrays, and
- (6) Expressed sequence tag (EST) analysis.

The above approaches have been used successfully to isolate differentially expressed genes in different model systems. However, each method has its own subtle (and sometimes not so subtle) characteristics which incur various advantages and disadvantages. Accordingly, it is the purpose of this review to clarify the mechanistic principles underlying the main differential expression methods and to highlight some of the broader considerations and implications of this very powerful and increasingly popular technique. Specifically, we will concentrate on the so-called 'open' systems, namely those which do not require any knowledge of gene sequences and, therefore, are useful for isolating unknown genes. Two 'closed' systems (those utilising previously identified gene sequences), EST analysis and the use of DNA arrays, will also be considered briefly for completeness. Whilst emphasis will often be placed on suppression PCR subtractive hybridization (SSH, the approach employed in this laboratory), it is the aim of the authors to highlight, wherever possible, those areas of common interest to those who use, or intend to use, differential gene expression analysis.

Differential cDNA library screening (DS)

Despite the development of multiple technological advances which have recently brought the field of gene expression profiling to the forefront of molecular analysis, recognition of the importance of differential gene expression and characterization of differentially expressed genes has existed for many years. One of the original approaches used to identify such genes was described 20 years ago by St John and Davis (1979). These authors developed a method, termed 'differential plaque filter

hybridization', which was used to isolate galactose-inducible DNA sequences from yeast. The theory is simple: a genomic DNA library is prepared from normal, unstimulated cells of the test organism/tissue and multiple filter replicas are prepared. These replica blots are probed with radioactively (or otherwise) labelled complex cDNA probes prepared from the control and test cell mRNA populations. Those mRNAs which are differentially expressed in the treated cell population will show a positive signal only on the filter probed with cDNA from the treated cells. Furthermore, labelled cDNA from different test conditions can be used to probe multiple blots, thereby enabling the identification of mRNAs which are only up-regulated under certain conditions. For example, St John and Davis (1979) screened replica filters with acetate-, glucose- and galactose-derived probes in order to obtain genes induced specifically by galactose metabolism. Although groundbreaking in its time this method is now considered insensitive and time-consuming, as up to 2 months are required to complete the identification of genes which are differentially expressed in the test population. In addition, there is no convenient way to check that the procedure has worked until the whole process has been completed.

Subtractive Hybridization (SH)

The developing concept of differential gene expression and the success of early approaches such as that described by St John and Davis (1979) soon gave rise to a search for more convenient methods of analysis. One of the first to be developed was SH, numerous variations of which have since been reported (see below). In general, this approach involves hybridization of mRNA/cDNA from one population (tester) to excess mRNA/cDNA from another (driver), followed by separation of the unhybridized tester fraction (differentially expressed) from the hybridized common sequences. This step has been achieved physically, chemically and through the use of selective polymerase chain reaction (PCR) techniques.

Physical separation

Original subtractive hybridization technology involved the physical separation of hybridized common species from unique single stranded species. Several methods of achieving this have been described, including hydroxyapatite chromatography (Sargent and Dawid 1983), avidin-biotin technology (Duguid and Dinauer 1990) and oligodT-latex separation (Hara *et al.* 1991). In the first approach, common mRNA species are removed by cDNA (from test cells)-mRNA (from control cells) subtractive hybridization followed by hydroxyapatite chromatography, as hydroxyapatite specifically adsorbs the cDNA-mRNA hybrids. The unabsorbed cDNA is then used either for the construction of a cDNA library of differentially expressed genes (Sargent and Dawid 1983, Schneider *et al.* 1988) or directly as a probe to screen a preselected library (Zimmerman *et al.* 1980, Davis *et al.* 1984, Hedrick *et al.* 1984). A schematic diagram of the procedure is shown in figure 1.

Less rigorous physical separation procedures coupled with sensitivity enhancing PCR steps were later developed as a means to overcome some of the problems encountered with the hydroxyapatite procedure. For example, Duguid and Dinauer (1990) described a method of subtraction utilizing biotin-affinity systems as a means to remove hybridized common sequences. In this process, both the control and tester mRNA populations are first converted to cDNA and an adaptor ('oligovector',

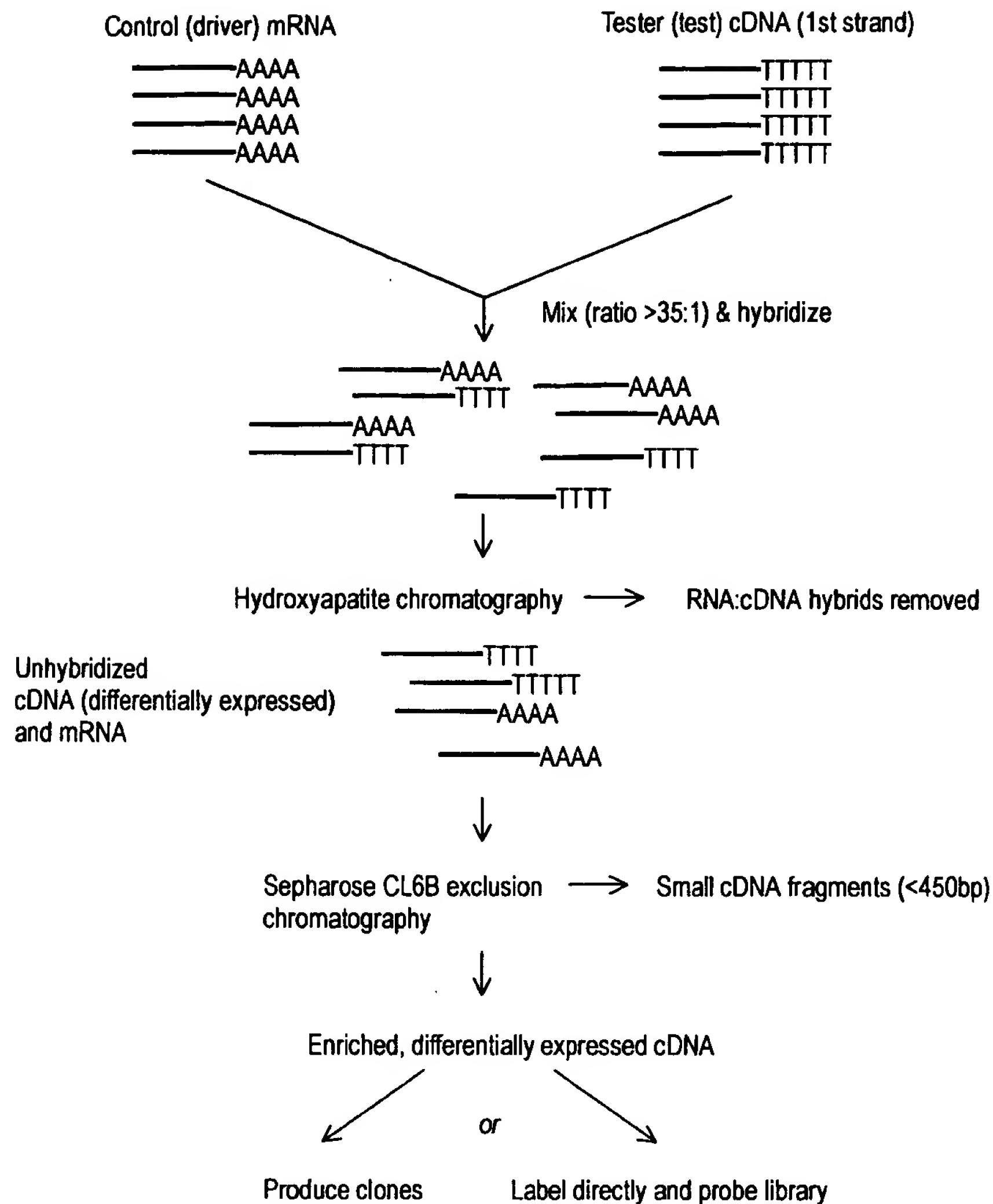


Figure 1. The hydroxyapatite method of subtractive hybridization. cDNA derived from the treated/alterd (tester) population is mixed with a large excess of mRNA from the control (driver) population. Following hybridization, mRNA-cDNA hybrids are removed by hydroxyapatite chromatography. The only cDNAs which remain are those which are differentially expressed in the treated/alterd population. In order to facilitate the recovery of full length clones, small cDNA fragments are removed by exclusion chromatography. The remaining cDNAs are then cloned into a vector for sequencing, or labelled and used directly to probe a library, as described by Sargent and Dawid (1983).

containing a restriction site) ligated to both sides. Both populations are then amplified by PCR, but the driver cDNA population is subsequently digested with the adaptor-containing restriction endonuclease. This serves to cleave the oligo-vector and reduce the amplification potential of the control population. The digested control population is then biotinylated and an excess mixed with tester cDNA. Following denaturation and hybridization, the mix is applied to a biocytin column (streptavidin may also be used) to remove the control population, including heteroduplexes formed by annealing of common sequences from the tester population. The procedure is repeated several times following the addition of fresh

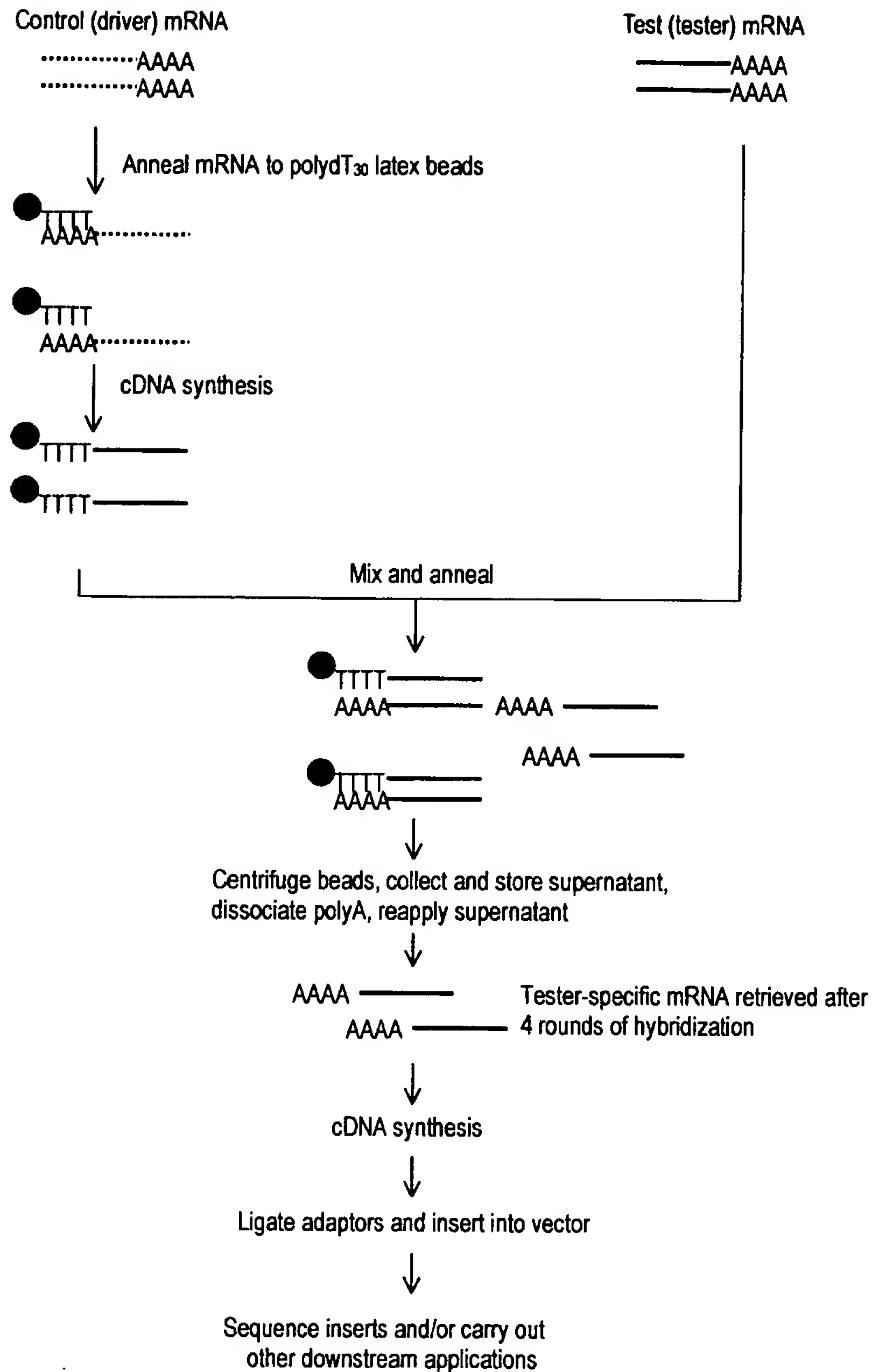


Figure 2. The use of oligodT₃₀ latex to perform subtractive hybridization. mRNA extracted from the control (driver) population is converted to anchored cDNA using polydT oligonucleotides attached to latex beads. mRNA from the treated/alterd (tester) population is repeatedly hybridized against an excess of the anchored driver cDNA. The final population of mRNA is tester specific and can be converted into cDNA for cloning and other downstream applications, as described by Hara *et al.* (1991).

control cDNA. In order to further enrich those species differentially expressed in the tester cDNA, the subtracted tester population is amplified by PCR following every second subtraction cycle. After six cycles of subtraction (three reamplification steps) the reaction mix is ligated into a vector for further analysis.

In a slightly different approach, Hara *et al.* (1991) utilized a method whereby oligo(dT₃₀) primers attached to a latex substrate are used to first capture mRNA extracted from the control population. Following 1st strand cDNA synthesis, the RNA strand of the heteroduplexes is removed by heat denaturation and centrifugation (the cDNA-oligotex-dT₃₀ forms a pellet and the supernatant is removed). A quantity of tester mRNA is then repeatedly hybridized to the immobilized control (driver) cDNA (which is present in 20-fold excess). After several rounds of hybridization the only mRNA molecules left in the tester mRNA population are those which are not found in the driver cDNA-oligotex-dT₃₀ population. These tester-specific mRNA species are then converted to cDNA and, following the addition of adaptor sequences, amplified by PCR. The PCR products are then ligated into a vector for further analysis using restriction sites incorporated into the PCR primers. A schematic illustration of this subtraction process is shown in figure 2.

However, all these methods utilising physical separation have been described as inefficient due to the requirement for large starting amounts of mRNA, significant loss of material during the separation process and a need for several rounds of hybridization. Hence, new methods of differential expression analysis have recently been designed to eliminate these problems.

Chemical Cross-Linking Subtraction (CCLS)

In this technique, originally described by Hampson *et al.* (1992), driver mRNA is mixed with tester cDNA (1st strand only) in a ratio of > 20:1. The common sequences form cDNA:mRNA hybrids, leaving the tester specific species as single stranded cDNA. Instead of physically separating these hybrids, they are inactivated chemically using 2,5 diaziridiny-1,4-benzoquinone (DZQ). Labelled probes are then synthesized from the remaining single stranded cDNA species (unreacted mRNA species remaining from the driver are not converted into probe material due to specificity of Sequenase T7 DNA polymerase used to make the probe) and used to screen a cDNA library made from the tester cell population. A schematic diagram of the system is shown in figure 3.

It has been shown that the differentially expressed sequences can be enriched at least 300-fold with one round of subtraction (Hampson *et al.* 1992), and that the technique should allow isolation of cDNAs derived from transcripts that are present at less than 50 copies per cell. This equates to genes at the low end of intermediate abundance (see table 1). The main advantages of the CCLS approach are that it is rapid, technically simple and also produces fewer false positives than other differential expression analysis methods. However, like the physical separation protocols, a major drawback with CCLS is the large amount of starting material required (at least 10 µg RNA). Consequently, the technique has recently been refined so that a renewable source of RNA can be generated. The degenerate random oligonucleotide primed (DROP) adaptation (Hampson *et al.* 1996, Hampson and Hampson 1997) uses random hexanucleotide sequences to prime solid phase-synthesized cDNA. Since each primer includes a T7 polymerase promotor sequence

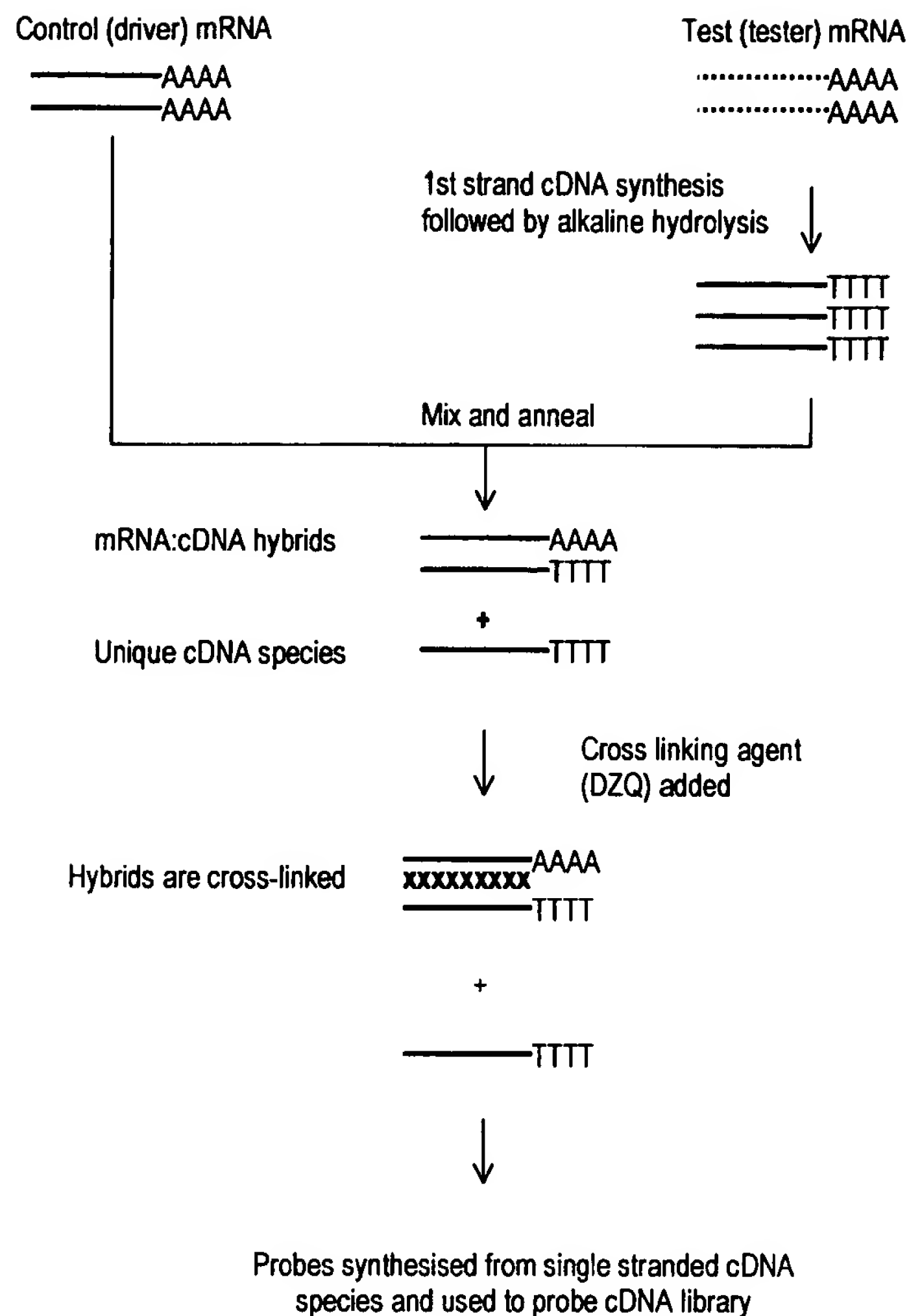


Figure 3. Chemical cross-linking subtraction. Excess driver mRNA is mixed with 1st strand tester cDNA. The common sequences form mRNA:cDNA hybrids which are cross linked with 2,5 diaziridinyl-1,4-benzoquinone (DZQ) and the remaining cDNA sequences are differentially expressed in the tester population. Probes are made from these sequences using Sequenase 2.0 DNA polymerase, which lacks reverse transcriptase activity and, therefore, does not react with the remaining mRNA molecules from the driver. The labelled probes are then used to screen a cDNA library for clones of differentially expressed sequences. Adapted from Walter *et al.* (1996), with permission.

Table 1. The abundance of mRNA species and classes in a typical mammalian cell.

mRNA class	Copies of each species/cell	No. of mRNA species in class	Mean % of each species in class	Mean mass (ng) of each species/ μ g total RNA
Abundant	12000	4	3.3	1.65
Intermediate	300	500	0.08	0.04
Rare	15	11000	0.004	0.002

Modified from Bertoli *et al.* (1995).

at the 5' end, the final pool of random cDNA fragments is a PCR-renewable cDNA population which is representative of the expressed gene pool and can be used to synthesize sense RNA for use as driver material. Furthermore, if the final pool of random cDNA fragments is reamplified using biotinylated T7 primer and random hexamer, the product can be captured with streptavidin beads and the antisense strand eluted for use as tester. Since both target and driver can be generated from the same DROP product, subtraction can be performed in both directions (i.e. for up- and down-regulated species) between two different DROP products.

Representational Difference Analysis (RDA)

RDA of cDNA (Hubank and Schatz 1994) is an extension of the technique originally applied to genomic DNA as a means of identifying differences between two complex genomes (Lisitsyn *et al.* 1993). It is a process of subtraction and amplification involving subtractive hybridization of the tester in the presence of excess driver. Sequences in the tester that have homologues in the driver are rendered unamplifiable, whereas those genes expressed only in the tester retain the ability to be amplified by PCR. The procedure is shown schematically in figure 4.

In essence, the driver and tester mRNA populations are first converted to cDNA and amplified by PCR following the ligation of an adaptor. The adaptors are then removed from both populations and a new (different) adaptor ligated to the amplified tester population only. Driver and tester populations are next melted and hybridized together in a ratio of 100:1. Following hybridization, only tester:tester homohybrids have 5' adaptors at each end of the DNA duplex and can, thus, be filled in at both 3' ends. Hence, only these molecules are amplified exponentially during the subsequent PCR step. Although tester:driver heterohybrids are present, they only amplify in a linear fashion, since the strand derived from the driver has no adaptor to which the primer can bind. Driver:driver heterohybrids have no adaptors and, therefore, are not amplified. Single stranded molecules are digested with mung bean nuclease before a further PCR-enrichment of the tester:tester homohybrids. The adaptors on the amplified tester population are then replaced and the whole process repeated a further two or three times using an increasing excess of driver (Hubank and Schatz used a tester:driver ratio of 1:400, 1:80000 and 1:800000 for the second, third and fourth hybridizations, respectively). Different adaptors are ligated to the tester between successive rounds of hybridization and amplification to prevent the accumulation of PCR products that might interfere with subsequent amplifications. The final display is a series of differentially expressed gene products easily observable on an ethidium bromide gel.

The main advantages of RDA are that it offers a reproducible and sensitive approach to the analysis of differentially expressed genes. Hubank and Schatz (1994) reported that they were able to isolate genes that were differentially expressed in substantially less than 1% of the cells from which the tester is derived. Perhaps the main drawback is that multiple rounds of ligation, hybridization, amplification and digestion are required. The procedure is, therefore, lengthier than many other differential display approaches and provides more opportunity for operator-induced error to occur. Although the generation of false positives has been noted, this has been solved to some degree by O'Neill and Sinclair (1997) through the use of HPLC-purified adaptors. These are free of the truncated adaptors which appear to be a major source of the false positive bands. A very similar technique to RDA, termed linker capture subtraction (LCS) was described by Yang and Sytowski (1996).

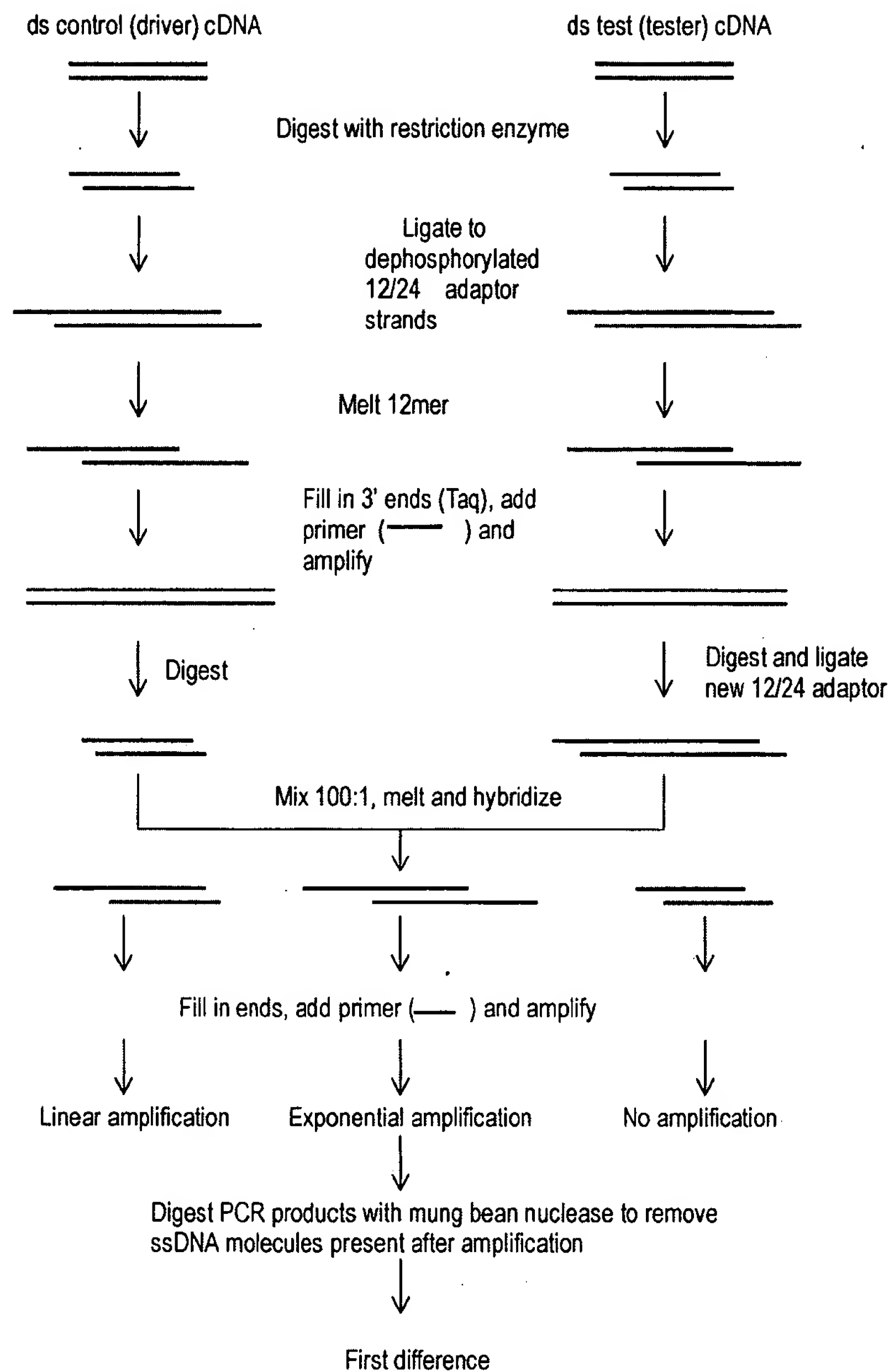


Figure 4. The representational difference analysis (RDA) technique. Driver and tester cDNA are digested with a 4-cutter restriction enzyme such as *DpnII*. The 1st set of 12/24 adaptor strands (oligonucleotides) are ligated to each other and the digested cDNA products. The 12mer is subsequently melted away and the 3' ends filled in using Taq DNA polymerase. Each cDNA population is then amplified using PCR, following which the 1st set of adaptors is removed with *DpnII*. A second set of 12/24 adaptor strands is then added to the amplified tester cDNA population, after which the tester is hybridized against a large excess of driver. The 12mer adaptors are melted and the 3' ends filled in as before. PCR is carried out with primers identical to the new 24mer adaptor. Thus, the only hybridization products which are exponentially amplified are those which are tester:tester combinations. Following PCR, ssDNA products are removed with mung bean nuclease, leaving the 'first difference product'. This is digested and a third set of 12/24 adaptors added before repeating the subtraction process from the hybridization stage. The process is repeated to the 3rd or 4th difference product, as described by Lisitsyn *et al.* (1993) and Hubank and Schatz (1994).

Suppression PCR Subtractive Hybridization (SSH)

The most recent adaptation of the SH approach to differential expression analysis was first described by Diatchenko *et al.* (1996) and Gurskaya *et al.* (1996). They reported that a 1000–5000 fold enrichment of rare cDNAs (equivalent to isolating mRNAs present at only a few copies per cell) can be obtained without the need for multiple hybridizations/subtractions. Instead of physical or chemical removal of the common sequences, a PCR-based suppression system is used (see figure 5).

In SSH, excess driver cDNA is added to two portions of the tester cDNA which have been ligated with different adaptors. A first round of hybridization serves to enrich differentially expressed genes and equalize rare and abundant messages. Equalization occurs since reannealing is more rapid for abundant molecules than for rarer molecules due to the second order kinetics of hybridization (James and Higgins 1985). The two primary hybridization mixes are then mixed together in the presence of excess driver and allowed to hybridize further. This step permits the annealing of single stranded complementary sequences which did not hybridize in the primary hybridization, and in doing so generates templates for PCR amplification. Although there are several possible combinations of the single stranded molecules present in the secondary hybridization mix, only one particular combination (differentially expressed in the tester cDNA composed of complimentary strands having different adaptors) can amplify exponentially.

Having obtained the final differential display, two options are available if cloning of cDNAs is desired. One is to transform the whole of the final PCR reaction into competent cells. Transformed colonies can then be isolated and their inserts characterized by sequencing, restriction analysis or PCR. Alternatively, the final PCR products can be resolved on a gel and the individual bands excised, reamplified and cloned. The first approach is technically simpler and less time consuming. However, ligation/transformation reactions are known to be biased towards the cloning of smaller molecules, and so the final population of clones will probably not contain a representative selection of the larger products. In addition, although equalization theoretically occurs, observations in this laboratory suggest that this is by no means perfectly accomplished. Consequently, some gene species are present in a higher number than others and this will be represented in the final population of clones. Thus, in order to obtain a substantial proportion of those gene species that actually demonstrate differential expression in the tester population, the number of clones that will have to be screened after this step may be substantial. The second approach is initially more time consuming and technically demanding. However, it would appear to offer better prospects for cloning larger and low abundance gel products. In addition, one can incorporate a screening step that differentiates different products of different sequences but of the same size (HA-staining, see later). In this way, a good idea of the final number of clones to be isolated and identified can be achieved.

An alternative (or even complementary) approach is to use the final differential display reaction to screen a cDNA library to isolate full length clones for further characterization, or a DNA array (see later) to quickly identify known genes. SSH has been used in this laboratory to begin characterization of the short-term gene expression profiles of enzyme-inducers such as phenobarbital (Rockett *et al.* 1997) and Wy-14,643 (Rockett *et al.* unpublished observations). The isolation of differentially expressed genes in this manner enables the construction of a fingerprint

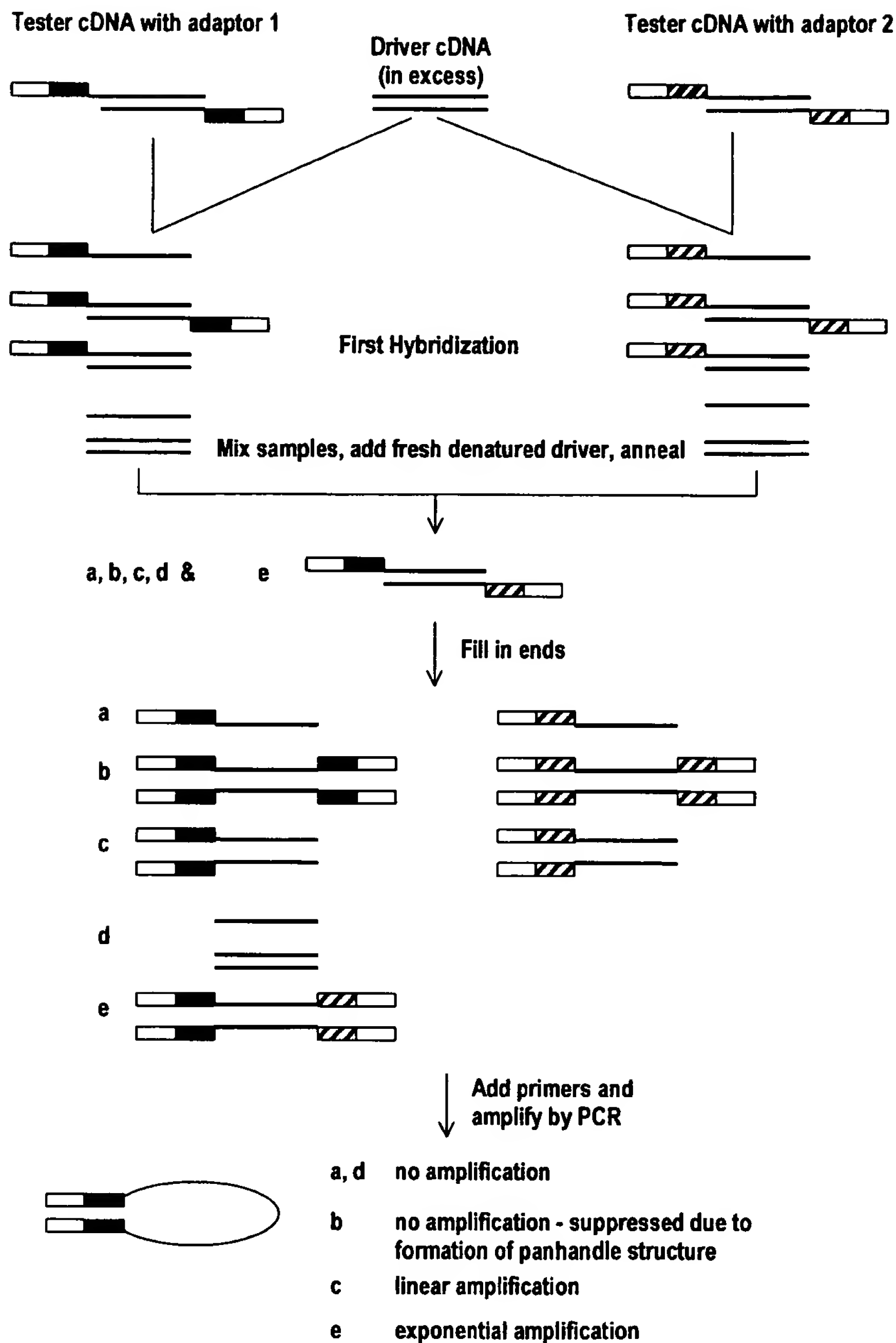


Figure 5. PCR-select cDNA subtraction. In the primary hybridization, an excess of driver cDNA is added to each tester cDNA population. The samples are heat denatured and allowed to hybridize for between 3 and 8 h. This serves two purposes: (1) to equalize rare and abundant molecules; and (2) to enrich for differentially expressed sequences—cDNAs that are not differentially expressed form type c molecules with the driver. In the secondary hybridization, the two primary hybridizations are mixed together without denaturing. Fresh denatured driver can also be added at this point to allow further enrichment of differentially expressed sequences. Type e molecules are formed in this secondary hybridization which are subsequently amplified using two rounds of PCR. The final products can be visualized on an agarose gel, labelled directly or cloned into a vector for downstream manipulation. As described by Diatchenko *et al.* (1996) and Gurskaya *et al.* (1996), with permission.

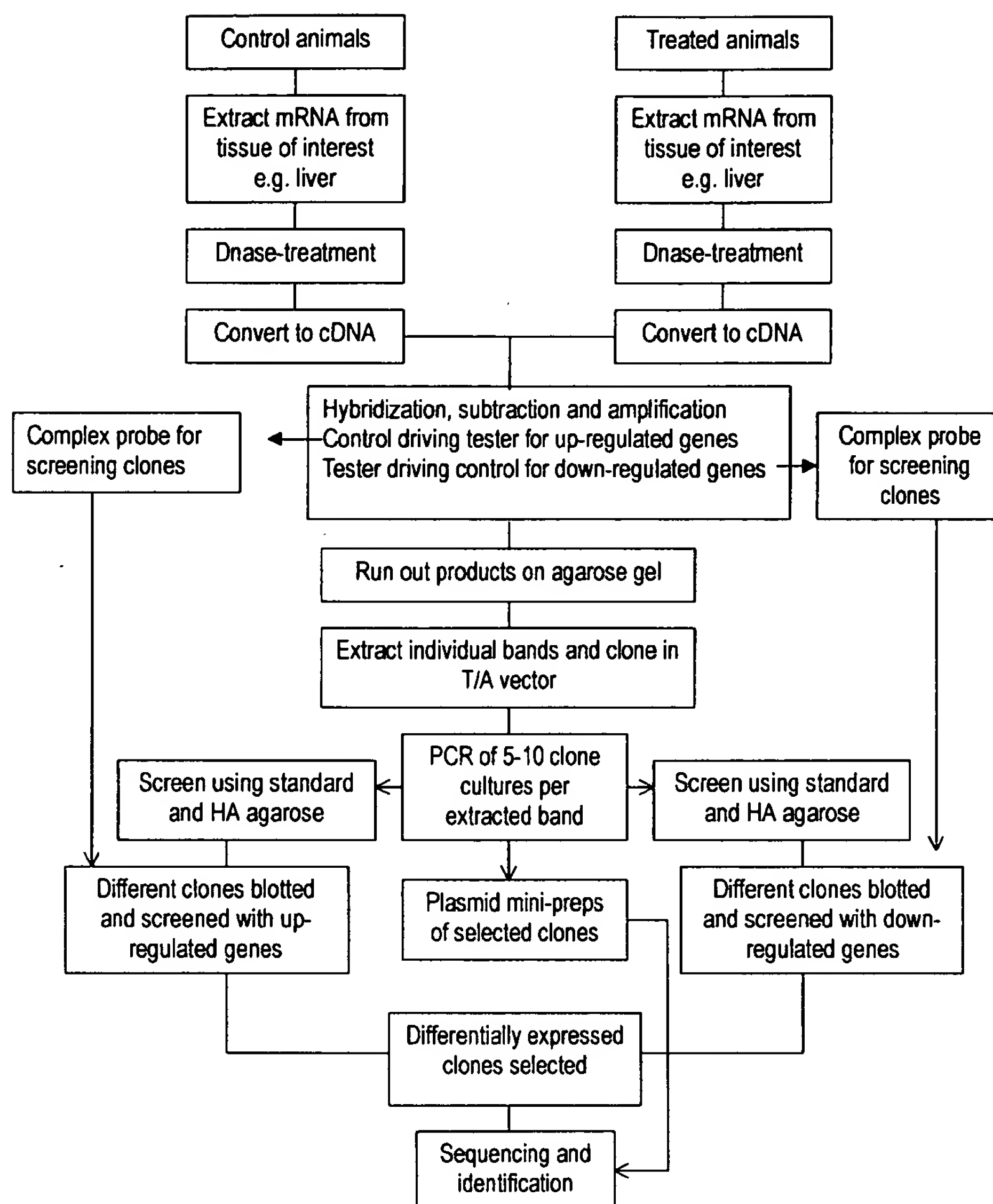


Figure 6. Flow diagram showing method used in this laboratory to isolate and identify clones of genes which are differentially expressed in rat liver following short term exposure to the enzyme inducers, phenobarbital and Wy-14,643.

of expressed genes which are unique to each compound and time/dose point. Such information could be useful in short-term characterization of the toxic potential of new compounds by comparing the gene-expression profiles they elicit with those produced by known inducers. Figure 6 shows a flow diagram of the method used to isolate, verify and clone differentially expressed genes, and figure 7 shows expression profiles obtained from a typical SSH experiment. Subsequent sub-cloning of the individual bands, sequencing and gene data base interrogation reveals many genes which are either up- or down-regulated by phenobarbital in the rat (tables 2 and 3).

One of the advantages in using the SSH approach is that no prior knowledge is required of which specific genes are up/down-regulated subsequent to xenobiotic

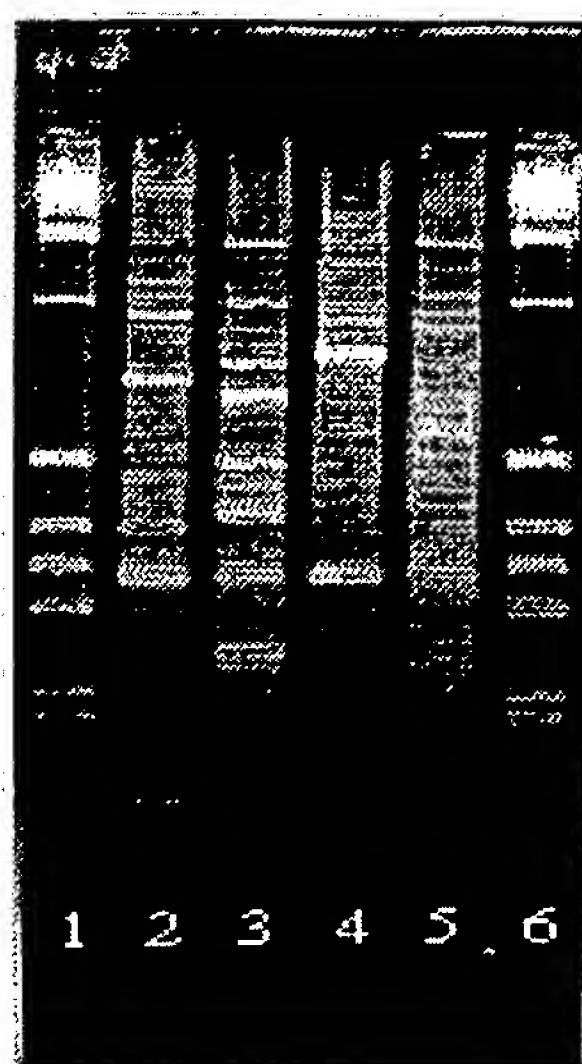


Figure 7. SSH display patterns obtained from rat liver following 3-day treatment with WY-14,643 or phenobarbital. mRNA extracted from control and treated livers was used to generate the differential displays using the PCR-Select cDNA subtraction kit (Clontech). Lane: 1—1kb ladder; 2—genes upregulated following Wy,14-643 treatment; 3—genes downregulated following Wy,14-643 treatment; 4—genes upregulated following phenobarbital treatment; 5—genes downregulated following phenobarbital treatment; 6—1kb ladder. Reproduced from Rockett *et al.* (1997), with permission.

exposure, and an almost complete complement of genes are obtained. For example, the peroxisome proliferator and non-genotoxic hepatocarcinogen Wy,14,643, up-regulates at least 28 genes and down-regulates at least 15 in the rat (a sensitive species) and produces 48 up- and 37 down-regulated genes in the guinea pig, a resistant species (Rockett, Swales, Esda and Gibson, unpublished observations). One of these genes, CD81, was up-regulated in the rat and down-regulated in the guinea pig following Wy-14,643 treatment. CD81 (alternatively named TAPA-1) is a widely expressed cell surface protein which is involved in a large number of cellular processes including adhesion, activation, proliferation and differentiation (Levy *et al.* 1998). Since all of these functions are altered to some extent in the phenomena of hepatomegaly and non-genotoxic hepatocarcinogenesis, it is intriguing, and probably mechanistically-relevant, that CD81 expression is differentially regulated in a resistant and susceptible species. However, the down-side of this approach is that the majority of genes can be sequenced and matched to database sequences, but the latter are predominantly expressed sequence tags or genes of completely unknown function, thus partially obscuring a realistic overall assessment of the critical genes of genuine biological interest. Notwithstanding the lack of complete functional identification of altered gene expression, such gene profiling studies essentially provides a 'molecular fingerprint' in response to xenobiotic challenge, thereby serving as a mechanistically-relevant platform for further detailed investigations.

Differential Display (DD)

Originally described as 'RNA fingerprinting by arbitrarily primed PCR' (Liang and Pardee 1992) this method is now more commonly referred to as 'differential

Table 2. Genes up-regulated in rat liver following 3-day exposure to phenobarbital.

Band number (approximate size in bp)	Highest sequence similarity	FASTA-EMBL gene identification
5 (1300)	93.5%	CYP2B1
7 (1000)	95.1%	Preproalbumin Serum albumin mRNA
8 (950)	98.3%	NCI-CGAP-Pr1 <i>H. sapiens</i> (EST)
10 (850)	95.7%	CYP2B1
11 (800)	Clone 1 94.9%	CYP2B1
	Clone 2 75.3%	CYP2B2
12 (750)	93.8%	TRPM-2 mRNA Sulfated glycoprotein
15 (600)	92.9%	Preproalbumin Serum albumin mRNA
16 (55)	Clone 1 95.2%	CYP2B1
	Clone 2 93.6%	Haptoglobin mRNA partial alpha
21 (350)	99.3%	18S, 5.8S & 28S rRNA

Bands 1–4, 6, 9, 13, 14, and 17–20 are shown to be false positives by dot blot analysis and, therefore, are not sequenced. Derived from Rockett *et al.* (1997). It should be noted that the above genes do not represent the complete spectrum of genes which are up-regulated in rat liver by phenobarbital, but simply represents the genes sequenced and identified to date.

Table 3. Genes down-regulated in rat liver following 3-day exposure to phenobarbital.

Band number (approximate size in bp)	Highest sequence similarity	FASTA-EMBL gene identification
1 (1500)	95.3%	3-oxoacyl-CoA thiolase
2 (1200)	92.3%	Hemopoxin mRNA
3 (1000)	91.7%	Alpha-2u-globulin mRNA
7 (700)	Clone 1 77.2%	<i>M. musculus</i> C1 inhibitor
	Clone 2 94.5%	Electron transfer flavoprotein
	Clone 3 91.0%	<i>M. musculus</i> Topoisomerase 1 (Topo 1)
8 (650)	Clone 1 86.9%	Soares 2NbMT <i>M. musculus</i> (EST)
	Clone 2 96.2%	Alpha-2u-globulin (s-type) mRNA
9 (600)	Clone 1 86.9%	Soares mouse NML <i>M. musculus</i> (EST)
	Clone 2 82.0%	Soares p3NMF 19.5 <i>M. musculus</i> (EST)
10 (550)	73.8%	Soares mouse NML <i>M. musculus</i> (EST)
11 (525)	95.7%	NCI-CGAP-Pr1 <i>H. sapiens</i> (EST)
12 (375)	100.0%	Ribosomal protein
13 (23)	Clone 1 97.2%	Soares mouse embryo NbME135 (EST)
	Clone 2 100.0%	Fibrinogen B-beta-chain
	Clone 3 100.0%	Apolipoprotein E gene
14 (170)	96.0%	Soares p3NMF19.5 <i>M. musculus</i> (EST)
15 (140)	97.3%	Stratagene mouse testis (EST)
Others: (300)	96.7%	<i>R. norvegicus</i> RASP 1 mRNA
(275)	93.1%	Soares mouse mammary gland (EST)

EST = Expressed sequence tag. Bands 4–6 were shown to be false positives by dot blot analysis and, therefore, were not sequenced. Derived from Rockett *et al.* (1997). It should be noted that the above genes do not represent the complete spectrum of genes which are down-regulated in rat liver by phenobarbital, but simply represents the genes sequenced and identified to date.

display' (DD). In this method, all the mRNA species in the control and treated cell populations are amplified in separate reactions using reverse transcriptase-PCR (RT-PCR). The products are then run side-by-side on sequencing gels. Those bands which are present in one display only, or which are much more intense in one

display compared to the other, are differentially expressed and may be recovered for further characterization. One advantage of this system is the speed with which it can be carried out—2 days to obtain a display and as little as a week to make and identify clones.

Two commonly used variations are based on different methods of priming the reverse transcription step (figure 8). One is to use an oligo dT with a 2-base 'anchor' at the 3'-end, e.g. 5' (dT₁₁)CA 3' (Liang and Pardee 1992). Alternatively, an arbitrary primer may be used for 1st strand cDNA synthesis (Welsh *et al.* 1992). This variant of RNA fingerprinting has also been called 'RAP' (RNA Arbitrarily Primed)-PCR. One advantage of this second approach is that PCR products may be derived from anywhere in the RNA, including open reading frames. In addition, it can be used for mRNAs that are not polyadenylated, such as many bacterial mRNAs (Wong and McClelland 1994). In both cases, following reverse transcription and denaturation, second strand cDNA synthesis is carried out with an arbitrary primer (*arbitrary* primers have a single base at each position, as compared to *random* primers, which contain a mixture of all four bases at each position). The resulting PCR, thus, produces a series of products which, depending on the system (primer length and composition, polymerase and gel system), usually includes 50–100 products per primer set (Band and Sager 1989). When a combination of different dT-anchors and arbitrary primers are used, almost all mRNA species from a cell can be amplified. When the cDNA products from two different populations are analysed side by side on a polyacrylamide gel, differences in expression can be identified and the appropriate bands recovered for cloning and further analysis.

Although DD is perhaps the most popular approach used today for identifying differentially expressed genes, it does suffer from several perceived disadvantages:

- (1) It may have a strong bias towards high copy number mRNAs (Bertioli *et al.* 1995), although this has been disputed (Wan *et al.* 1996) and the isolation of very low abundance genes may be achieved in certain circumstances (Guimeraes *et al.* 1995a).
- (2) The cDNAs obtained often only represent the extreme 3' end of the mRNA (often the 3'-untranslated region), although this may not always be the case (Guimeraes *et al.* 1995a). Since the 3' end is often not included in Genbank and shows variation between organisms, cDNAs identified by DD cannot always be matched with their genes, even if they have been identified.
- (3) The pattern of differential expression seen on the display often cannot be reproduced on Northern blots, with false positives arising in up to 70% of cases (Sun *et al.* 1994). Some adaptations have been shown to reduce false positives, including the use of two reverse transcriptases (Sung and Denman 1997), comparison of uninduced and induced cells over a time course (Burn *et al.* 1994) and comparison of DDPCR-products from two uninduced and two induced lines (Sompayrac *et al.* 1995). The latter authors also reported that the use of cytoplasmic RNA rather than total RNA reduces false positives arising from nuclear RNA that is not transported to the cytoplasm.

Further details of the background, strengths and weaknesses of the DD technique can be obtained from a review by McClelland *et al.* (1996) and from articles by Liang *et al.* (1995) and Wan *et al.* (1996).

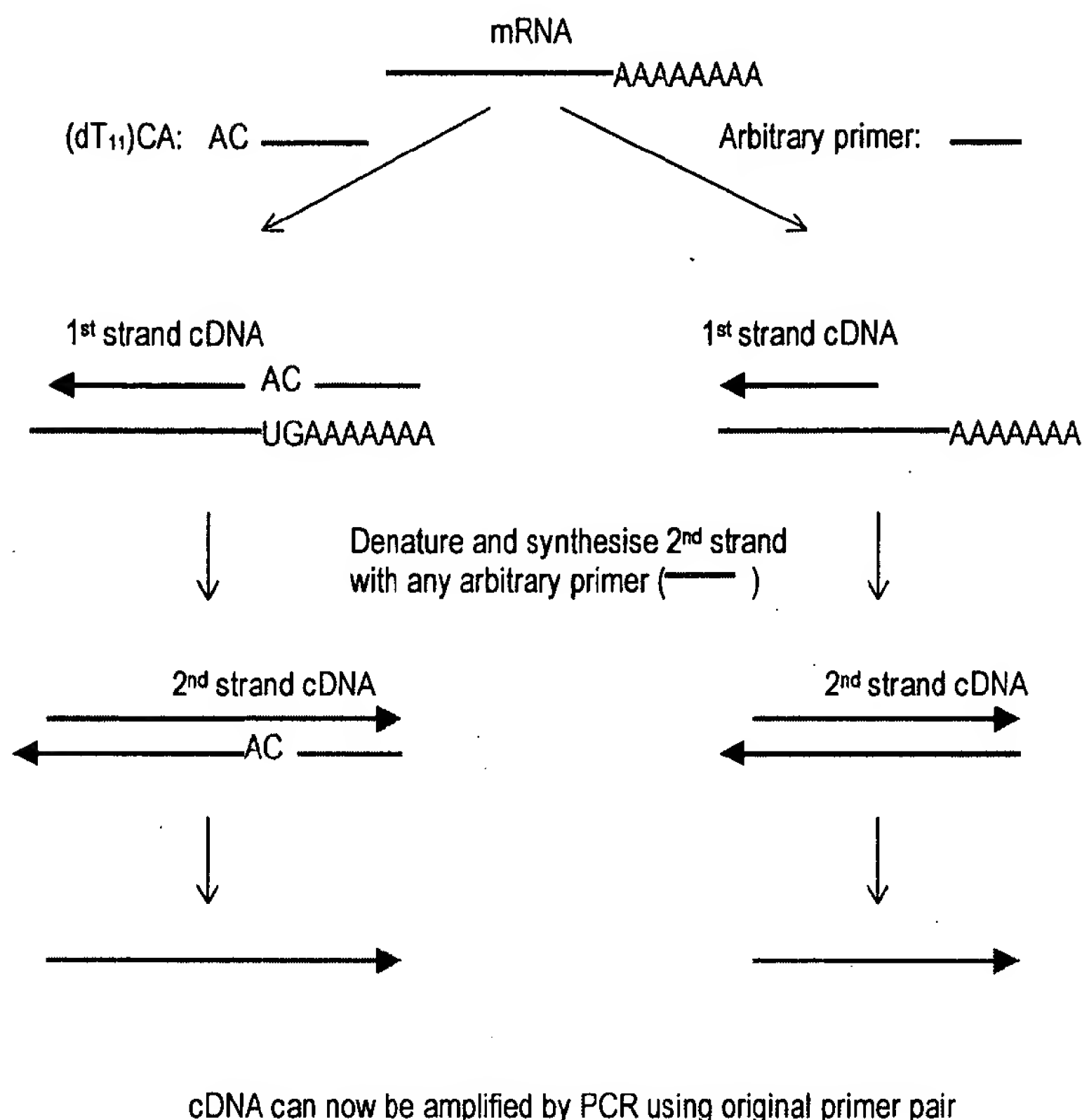


Figure 8. Two approaches to differential display (DD) analysis. 1st strand synthesis can be carried out either with a polydT₁₁NN primer (where N = G, C or A) or with an arbitrary primer. The use of different combinations of G, C and A to anchor the first strand polydT primer enables the priming of the majority of polyadenylated mRNAs. Arbitrary primers may hybridize at none, one or more places along the length of the mRNA, allowing 1st strand cDNA synthesis to occur at none, one or more points in the same gene. In both cases, 2nd strand synthesis is carried out with an arbitrary primer. Since these arbitrary primers for the 2nd strand may also hybridize to the 1st strand cDNA in a number of different places, several different 2nd strand products may be obtained from one binding point of the 1st strand primer. Following 2nd strand synthesis, the original set of primers is used to amplify the second strand products, with the result that numerous gene sequences are amplified.

Restriction endonuclease-facilitated analysis of gene expression

Serial Analysis of Gene Expression (SAGE)

A more recent development in the field of differential display is SAGE analysis (Velculescu *et al.* 1995). This method uses a different approach to those discussed so far and is based on two principles. Firstly, in more than 95% of cases, short nucleotide sequences ('tags') of only nine or 10 base pairs provide sufficient information to identify their gene of origin. Secondly, concatenation (linking together in a series) of these tags allows sequencing of multiple cDNAs within a single clone. Figure 9 shows a schematic representation of the SAGE process. In this procedure, double stranded cDNA from the test cells is synthesized with a biotinylated polydT primer. Following digestion with a commonly cutting (4bp recognition sequence) restriction enzyme ('anchoring enzyme'), the 3' ends of the cDNA population are captured with streptavidin beads. The captured population is

split into two and different adaptors ligated to the 5' ends of each group. Incorporated into the adaptors is a recognition sequence for a type IIS restriction enzyme—one which cuts DNA at a defined distance (< 20 bp) from its recognition sequence. Hence, following digestion of each captured cDNA population with the IIS enzyme, the adaptors plus a short piece of the captured cDNA are released. The two populations are then ligated and the products amplified. The amplified products are cleaved with the original anchoring enzyme, religated (concatomers are formed in the process) and cloned. The advantage of this system is that hundreds of gene tags can be identified by sequencing only a few clones. Furthermore, the number of times a given transcript is identified is a quantitative measurement of that gene's abundance in the original population, a feature which facilitates identification of differentially expressed genes in different cell populations.

Some disadvantages of SAGE analysis include the technical difficulty of the method, a large amount of accurate sequencing is required, biased towards abundant mRNAs, has not been validated in the pharmaco/toxicogenomic setting and has only been used to examine well known tissue differences to date.

Gene Expression Fingerprinting (GEF)

A different capture/restriction digest approach for isolating differentially expressed genes has been described by Ivanova and Belyavsky (1995). In this method, RNA is converted to cDNA using biotinylated oligo(dT) primers. The cDNA population is then digested with a specific endonuclease and captured with magnetic streptavidin microbeads to facilitate removal of the unwanted 5' digestion products. The use of restricted 3'-ends alone serves to reduce the complexity of the cDNA fragment pool and helps to ensure that each RNA species is represented by not more than one restriction product. An adaptor is ligated to facilitate subsequent amplification of the captured population. PCR is carried out with one adaptor-specific and one biotinylated polydT primer. The reamplified population is recaptured and the non-biotinylated strands removed by alkaline dissociation. The non-biotinylated strand is then resynthesized using a different adaptor-specific primer in the presence of a radiolabelled dNTP. The labelled immobilized 3' cDNA ends are next sequentially treated with a series of different restriction endonucleases and the products from each digestion analysed by PAGE. The result is a fingerprint composed of a number of ladders (equal to the number of sequential digests used). By comparing test versus control fingerprints, it is possible to identify differentially expressed products which can then be isolated from the gel and cloned. The advantages of this procedure are that it is very robust and reproducible, and the authors estimate that 80–93% of cDNA molecules are involved in the final fingerprint. The disadvantage is that polyacrylamide gels can rarely resolve more than 300–400 bands, which compares poorly to the 1000 or more which are estimated to be produced in an average experiment. The use of 2-D gels such as those described by Uitterlinden *et al.* (1989) and Hatada *et al.* (1991) may help to overcome this problem.

A similar method for displaying restriction endonuclease fragments was later described by Prashar and Weissman (1996). However, instead of sequential digestion of the immobilized 3'-terminal cDNA fragments, these authors simply compared the profiles of the control and treated populations without further manipulation.

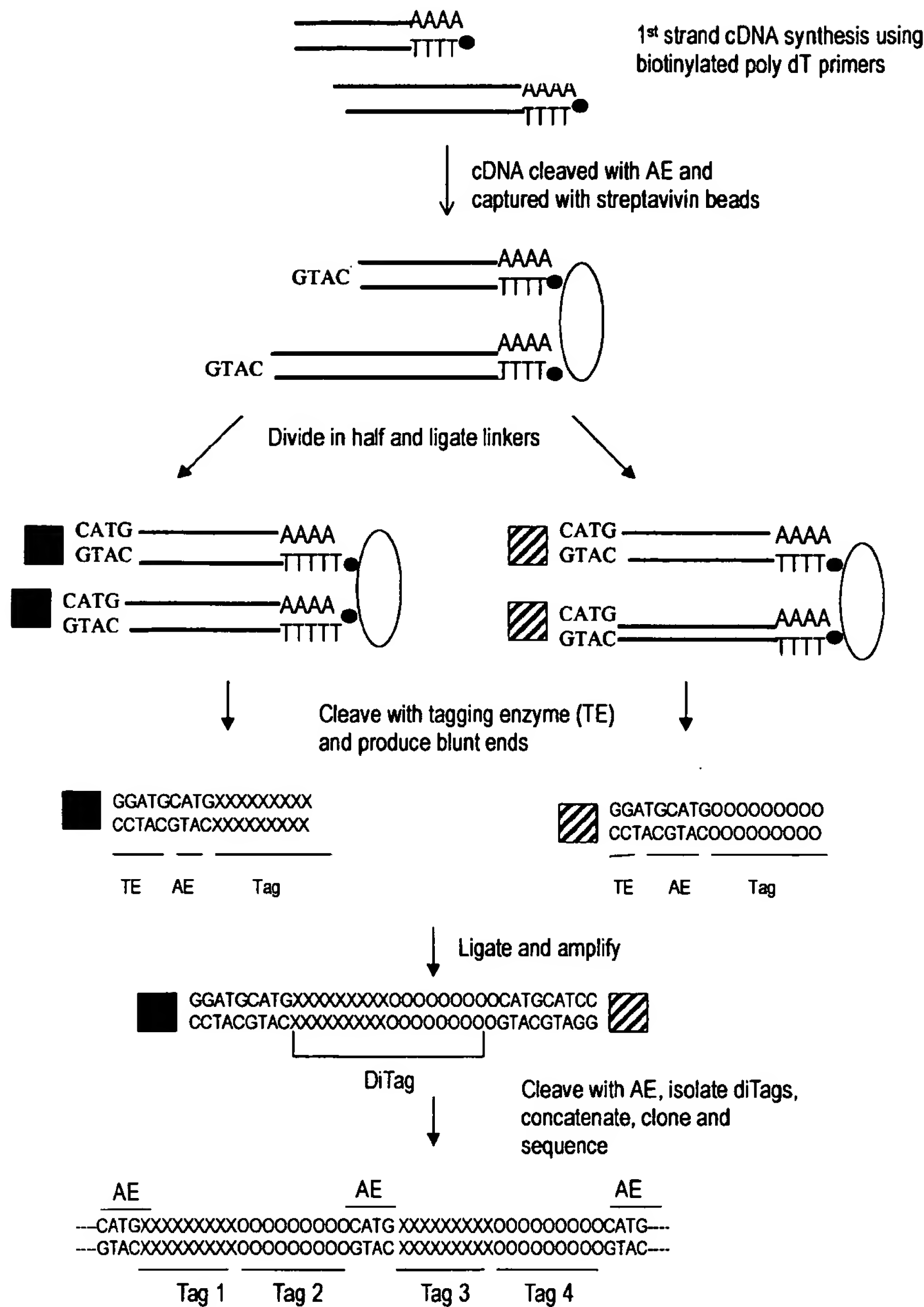


Figure 9. Serial analysis of gene expression (SAGE) analysis. cDNA is cleaved with an anchoring enzyme (AE) and the 3' ends captured using streptavidin beads. The cDNA pool is divided in half and each portion ligated to a different linker, each containing a type IIS restriction site (tagging enzyme, TE). Restriction with the type IIS enzyme releases the linker plus a short length of cDNA (XXXXX and OOOOO indicate nucleotides of different tags). The two pools of tags are then ligated and amplified using linker-specific primers. Following PCR, the products are cleaved with the AE and the diTags isolated from the linkers using PAGE. The diTags are then ligated (during which process, concatenization occurs) and cloned into a vector of choice for sequencing. After Velculescu *et al.* (1995), with permission.

DNA arrays

'Open' differential display systems are cumbersome in that it takes a great deal of time to extract and identify candidate genes and then confirm that they are indeed up- or down-regulated in the treated compared to the control tissue. Normally, the latter process is carried out using Northern blotting or RT-PCR. Even so, each of the aforementioned steps produce a bottleneck to the ultimate goal of rapid analysis of gene expression. These problems will likely be addressed by the development of so-called DNA arrays (e.g. Gress *et al.* 1992, Zhao *et al.* 1995, Schena *et al.* 1996), the introduction of which has signalled the next era in differential gene expression analysis. DNA arrays consist of a gridded membrane or glass 'chips' containing hundreds or thousands of DNA spots, each consisting of multiple copies of part of a known gene. The genes are often selected based on previously proven involvement in oncogenesis, cell cycling, DNA repair, development and other cellular processes. They are usually chosen to be as specific as possible for each gene and animal species. Human and mouse arrays are already commercially available and a few companies will construct a personalized array to order, for example Clontech Laboratories and Research Genetics Inc. The technique is rapid in that hundreds or even thousands of genes can be spotted on a single array, and that mRNA/cDNA from the test populations can be labelled and used directly as probe. When analysed with appropriate hardware and software, arrays offer a rapid and quantitative means to assess differences in gene expression between two cell populations. Of course, there can only be identification and quantitation of those genes which are in the array (hence the term 'closed' system). Therefore, one approach to elucidating the molecular mechanisms involved in a particular disease/development system may be to combine an open and closed system—a DNA array to directly identify and quantitate the expression of known genes in mRNA populations, and an open system such as SSH to isolate unknown genes which are differentially expressed.

One of the main advantages of DNA arrays is the huge number of gene fragments which can be put on a membrane—some companies have reported gridding up to 60 000 spots on a single glass 'chip' (microscope slide). These high density chip-based micro-arrays will probably become available as mass-produced off-the-shelf items in the near future. This should facilitate the more rapid determination of differential expression in time and dose-response experiments. Aside from their high cost and the technical complexities involved in producing and probing DNA arrays, the main problem which remains, especially with the newer micro-array (gene-chip) technologies, is that results are often not wholly reproducible between arrays. However, this problem is being addressed and should be resolved within the next few years.

EST databases as a means to identify differentially expressed genes

Expressed sequence tags (ESTs) are partial sequences of clones obtained from cDNA libraries. Even though most ESTs have no formal identity (putative identification is the best to be hoped for), they have proven to be a rapid and efficient means of discovering new genes and can be used to generate profiles of gene-expression in specific cells. Since they were first described by Adams *et al.* (1991), there has been a huge explosion in EST production and it is estimated that there are now well over a million such sequences in the public domain, representing over half

of all human genes (Hillier *et al.* 1996). This large number of freely available sequences (both sequence information and clones are normally available royalty-free from the originators) has enabled the development of a new approach towards differential gene expression analysis as described by Vasmatazis *et al.* (1998). The approach is simple in theory: EST databases are first searched for genes that have a number of related EST sequences from the target tissue of choice, but none or few from non-target tissue libraries. Programmes to assist in the assembly of such sets of overlapping data may be developed in-house or obtained privately or from the internet. For example, the Institute for Genomic Research (TIGR, found at <http://www.tigr.org>) provides many software tools free of charge to the scientific community. Included amongst these is the TIGR assembler (Sutton *et al.* 1995), a tool for the assembly of large sets of overlapping data such as ESTs, bacterial artificial chromosomes (BAC)s, or small genomes. Candidate EST clones representing different genes are then analysed using RNA blot methods for size and tissue specificity and, if required, used as probes to isolate and identify the full length cDNA clone for further characterization. In practice however, the method is rather more involved, requiring bioinformatic and computer analysis coupled with confirmatory molecular studies. Vasmatazis *et al.* (1998) have described several problems in this fledgling approach, such as separating highly homologous sequences derived from different genes and an overemphasis of specificity for some EST sequences. However, since these problems will largely be addressed by the development of more suitable computer algorithms and an increased completeness of the EST database, it is likely that this approach to identifying differentially expressed genes may enjoy more patronage in the future.

Problems and potential of differential expression techniques

The holistic or single cell approach?

When working with *in vivo* models of differential expression, one of the first issues to consider must be the presence of multiple cell types in any given specimen. For example, a liver sample is likely to contain not only hepatocytes, but also (potentially) Ito cells, bile ductule cells, endothelial cells, various immune cells (e.g. lymphocytes, macrophages and Kupffer cells) and fibroblasts. Other tissues will each have their own distinctive cell populations. Also, in the case of neoplastic tissue, there are almost always normal, hyperplastic and/or dysplastic cells present in a sample. One must, therefore, be aware that genes obtained from a differential display experiment performed on an animal tissue model may not necessarily arise exclusively from the intended 'target' cells, e.g. hepatocytes/neoplastic cells. If appropriate, further analyses using immunohistochemistry, *in situ* hybridization or *in situ* RT-PCR should be used to confirm which cell types are expressing the gene(s) of interest. This problem is probably most acute for those studying the differential expression of genes in the development of different cell types, where there is a need to examine homologous cell populations. The problem is now being addressed at the National Cancer Institute (Bethesda, MD, USA) where new microdissection techniques have been employed to assist in their gene analysis programme, the Cancer Genome Anatomy Project (CGAP) (For more information see web site: <http://www.ncbi.nlm.nih.gov/ncicgap/intro.html>). There are also separation techniques available that utilise cell-specific antigens as a means to isolate target cells,

e.g. fluorescence activated cell sorting (FACS) (Dunbar *et al.* 1998, Kas-Deelen *et al.* 1998) and magnetic bead technology (Richard *et al.* 1998, Rogler *et al.* 1998).

However, those taking a holistic approach may consider this issue unimportant. There is an equally appropriate view that all those genes showing altered expression within a compromised tissue should be taken into consideration. After all, since all tissues are complex mixes of different, interacting cell types which intimately regulate each other's growth and development, it is clear that each cell type could in some way contribute (positively or negatively) towards the molecular mechanisms which lie behind responses to external stimuli or neoplastic growth. It is perhaps then more informative to carry out differential display experiments using *in vivo* as opposed to *in vitro* models, where uniform populations of identical cells probably represent a partial, skewed or even inaccurate picture of the molecular changes that occur.

The incidence and possible implications of inter-individual biological variation should be considered in any approach where whole animal models are being used. It is clear that individuals (humans and animals) respond in different ways to identical stimuli. One of the best characterized examples is the debrisoquine oxidation polymorphism, which is mediated by cytochrome CYP2D6 and determines the pharmacokinetics of many commonly prescribed drugs (Lennard 1993, Meyer and Zanger 1997). The reasons for such differences are varied and complex, but allelic variations, regulatory region polymorphisms and even physical and mental health can all contribute to observed differences in individual responses. Careful thought should, therefore, be given to the specific objectives of the study and to the possible value of pooling starting material (tissue/mRNA). The effect of this can be beneficial through the ironing out of exaggerated responses and unimportant minor fluctuations of (mechanistically) irrelevant genes in individual animals, thus providing a clearer overall picture of the general molecular mechanisms of the response. However, at the same time such minor variations may be of utmost importance in deciding the ability of individual animals to succumb to or resist the effects of a given chemical/disease.

How efficient are differential expression techniques at recovering a high percentage of differentially expressed genes?

A number of groups have produced experimental data suggesting that mammalian cells produce between 8000–15 000 different mRNA species at any one time (Mechler and Rabbitts 1981, Hedrick *et al.* 1984, Bravo 1990), although figures as high as 20–30 000 have also been quoted (Axel *et al.* 1976). Hedrick *et al.* (1984) provided evidence suggesting that the majority of these belong to the rare abundance class. A breakdown of this abundance distribution is shown in table 1.

When the results of differential display experiments have been compared with data obtained previously using other methods, it is apparent that not all differentially expressed mRNAs are represented in the final display. In particular, rare messages (which, importantly, often include regulatory proteins) are not easily recovered using differential display systems. This is a major shortcoming, as the majority of mRNA species exist at levels of less than 0.005% of the total population (table 1). Bertoli *et al.* (1995) examined the efficiency of DD templates (heterogeneous mRNA populations) for recovering rare messages and were unable to detect mRNA

species present at less than 1.2% of the total mRNA population—equivalent to an intermediate or abundant species. Interestingly, when simple model systems (single target only) were used instead of a heterogeneous mRNA population, the same primers could detect levels of target mRNA down to 10000× smaller. These results are probably best explained by competition for substrates from the many PCR products produced in a DD reaction.

The numbers of differentially expressed mRNAs reported in the literature using various model systems provides further evidence that many differentially expressed mRNAs are not recovered. For example, DeRisi *et al.* (1997) used DNA array technology to examine gene expression in yeast following exhaustion of sugar in the medium, and found that more than 1700 genes showed a change in expression of at least 2-fold. In light of such a finding, it would not be unreasonable to suggest that of the 8000–15 000 different mRNA species produced by any given mammalian cell, up to 1000 or more may show altered expression following chemical stimulation. Whilst this may be an extreme figure, it is known that at least 100 genes are activated/upregulated in Jurkat (T-) cells following IL-2 stimulation (Ullman *et al.* 1990). In addition, Wan *et al.* (1996) estimated that interferon- γ -stimulated HeLa cells differentially express up to 433 genes (assuming 24000 distinct mRNAs expressed by the cells). However, there have been few publications documenting anywhere near the recovery of these numbers. For example, in using DD to compare normal and regenerating mouse liver, Bauer *et al.* (1993) found only 70 of 38000 total bands to be different. Of these, 50% (35 genes) were shown to correspond to differentially expressed bands. Chen *et al.* (1996) reported 10 genes upregulated in female rat liver following ethinyl estradiol treatment. McKenzie and Drake (1997) identified 14 different gene products whose expression was altered by phorbol myristate acetate (PMA, a tumour promoter agent) stimulation of a human myelomonocytic cell line. Kilty and Vickers (1997) identified 10 different gene products whose expression was upregulated in the peripheral blood leukocytes of allergic disease sufferers. Linskens *et al.* (1995) found 23 genes differentially expressed between young and senescent fibroblasts. Techniques other than DD have also provided an apparent paucity of differentially expressed genes. Using SH for example, Cao *et al.* (1997) found 15 genes differentially expressed in colorectal cancer compared to normal mucosal epithelium. Fitzpatrick *et al.* (1995) isolated 17 genes upregulated in rat liver following treatment with the peroxisome proliferator, clofibrate; Philips *et al.* (1990) isolated 12 cDNA clones which were upregulated in highly metastatic mammary adenocarcinoma cell lines compared to poorly metastatic ones. Prashar and Weissman (1996) used 3' restriction fragment analysis and identified approximately 40 genes showing altered expression within 4 h of activation of Jurkat T-cells. Groenink and Leegwater (1996) analysed 27 gene fragments isolated using SSH of delayed early response phase of liver regeneration and found only 12 to be upregulated.

In the laboratory, SSH was used to isolate up to 70 candidate genes which appear to show altered expression in guinea pig liver following short-term treatment with the peroxisome proliferator, WY-14,643 (Rockett, Swales, Esdaile and Gibson, unpublished observations). However, these findings have still to be confirmed by analysis of the extracted tissue mRNA for differential expression of these sequences.

Whilst the latest differential display technologies are purported to include design and experimental modifications to overcome this lack of efficiency (in both the total number of differentially expressed genes recovered and the percentage that are true

positives), it is still not clear if such adaptations are practically effective—proving efficiency by spiking with a known amount of limited numbers of artificial construct(s) is one thing, but isolating a high percentage of the rare messages already present in an mRNA population is another. Of course, some models will genuinely produce only a small number of differentially expressed genes. In addition, there are also technical problems that can reduce efficiency. For example, mRNAs may have an unusual primary structure that effectively prevents their amplification by PCR-based systems. In addition, it is known that under certain circumstances not all mRNAs have 3' polyA sites. For example, during *Xenopus* development, deadenylation is used as a means to stabilize RNAs (Voeltz and Steitz 1998), whilst preferential deadenylation may play a role in regulating Hsp70 (and perhaps, therefore, other stress protein) expression in *Drosophila* (Dellavalle *et al.* 1994). The presence of deadenylated mRNAs would clearly reduce the efficiency of systems utilizing a polydT reverse transcription step. The efficiency of any system also depends on the quality of the starting material. All differential display techniques use mRNA as their target material. However, it is difficult to isolate mRNA that is completely free of ribosomal RNA. Even if polydT primers are used to prime first strand cDNA synthesis, ribosomal RNA is often transcribed to some degree (Clontech PCR-Select cDNA Subtraction kit user manual). It has been shown, at least in the case of SSH, that a high rRNA:mRNA ratio can lead to inefficient subtractive hybridization (Clontech PCR-Select cDNA Subtraction kit user manual), and there is no reason to suppose that it will not do likewise in other SH approaches. Finally, those techniques that utilise a presubtraction amplification step (e.g. RDA) may present a skewed representation since some sequences amplify better than others.

Of course, probably the most important consideration is the temporal factor. It is clear that any given differential display experiment can only interrogate a cell at one point in time. It may well be that a high percentage of the genes showing altered expression at that time are obtained. However, given that disease processes and responses to environmental stimuli involve dynamic cascades of signalling, regulation, production and action, it is clear that all those genes which are switched on/off at different times will not be recovered and, therefore, vital information may well be missed. It is, therefore, imperative to obtain as much information about the model system beforehand as possible, from which a strategy can be derived for targeting specific time points or events that are of particular interest to the investigator. One way of getting round this problem of single time point analysis is to conduct the experiment over a suitable time course which, of course, adds substantially to the amount of work involved.

How sensitive are differential expression technologies?

There has been little published data that addresses the issue of how large the change in expression must be for it to permit isolation of the gene in question with the various differential expression technologies. Although the isolation of genes whose expression is changed as little as 1.5-fold has been reported using SSH (Groenink and Leegwater 1996), it appears that those demonstrating a change in excess of 5-fold are more likely to be picked up. Thus, there is a 'grey zone' in between where small changes could fade in and out of isolation between

experiments and animals. DD, on the other hand, is not subject to this grey zone since, unlike SH approaches, it does not amplify the difference in expression between two samples. Wan *et al.* (1996) reported that differences in expression of twofold or more are detectable using DD.

Resolution and visualization of differential expression products

It seems highly improbable with current technology that a gel system could be developed that is able to resolve all gene species showing altered expression in any given test system (be it SH- or DD-based). Polyacrylamide gel electrophoresis (PAGE) can resolve size differences down to 0.2% (Sambrook *et al.* 1989) and are used as standard in DD experiments. Even so, it is clear that a complex series of gene products such as those seen in a DD will contain unresolvable components. Thus, what appears to be one band in a gel may in fact turn out to be several. Indeed, it has been well documented (Mathieu-Daude *et al.* 1996, Smith *et al.* 1997) that a single band extracted from a DD often represents a composite of heterogeneous products, and the same has been found for SSH displays in this laboratory (Rockett *et al.* 1997). One possible solution was offered by Mathieu-Daude *et al.* (1996), who extracted and reamplified candidate bands from a DD display and used single strand conformation polymorphism (SSCP) analysis to confirm which components represented the truly differentially expressed product.

Many scientists often try to avoid the use of PAGE where possible because it is technically more demanding than agarose gel electrophoresis (AGE). Unfortunately, high resolution agarose gels such as Metaphor (FMC, Lichfield, UK) and AquaPor HR (National Diagnostics, Hesse, UK), whilst easier to prepare and manipulate than PAGE, can only separate DNA sequences which differ in size by around 1.5–2% (15–20 base pairs for a 1Kb fragment). Thus, SSH, RDA or other such products which differ in size by less than this amount are normally not resolvable. However, a simple technique does in fact exist for increasing the resolving power of AGE—the inclusion of HA-red (10-phenyl neutral red-PEG ligand) or HA-yellow (bisbenzamide-PEG ligand) (Hanse Analytik GmbH, Bremen, Germany) in a gel separates identical or closely sized products on base content. Specifically, HA-red and -yellow selectively bind to GC and AT DNA motifs, respectively (Wawer *et al.* 1995, Hanse Analytik 1997, personal communication). Since both HA-stains possess an overall positive charge, they migrate towards the cathode when an electric field is applied. This is in direct opposition to DNA, which is negatively charged and, therefore, migrates towards the anode. Thus, if two DNA clones are identical in size (as perceived on a standard high resolution agarose gel), but differ in AT/GC content, inclusion of a HA-dye in the gel will effectively retard the migration of one of the sequences compared to the other, effectively making it apparently larger and, thus, providing a means of differentiating between the two. The use of HA-red has been shown to resolve sequences with an AT variation of less than 1% (Wawer *et al.* 1995), whilst Hanse Analytik have reported that HA staining is so sensitive that in one case it was used to distinguish two 567bp sequences which differed by only a single point mutation (Hanse Analytik 1996, personal communication). Therefore, if one wishes to check whether all the clones produced from a specific band in a differential display experiment are derived from the same gene species, a small amount of reamplified or digested clone can be run on a standard high resolution gel, and a second aliquot

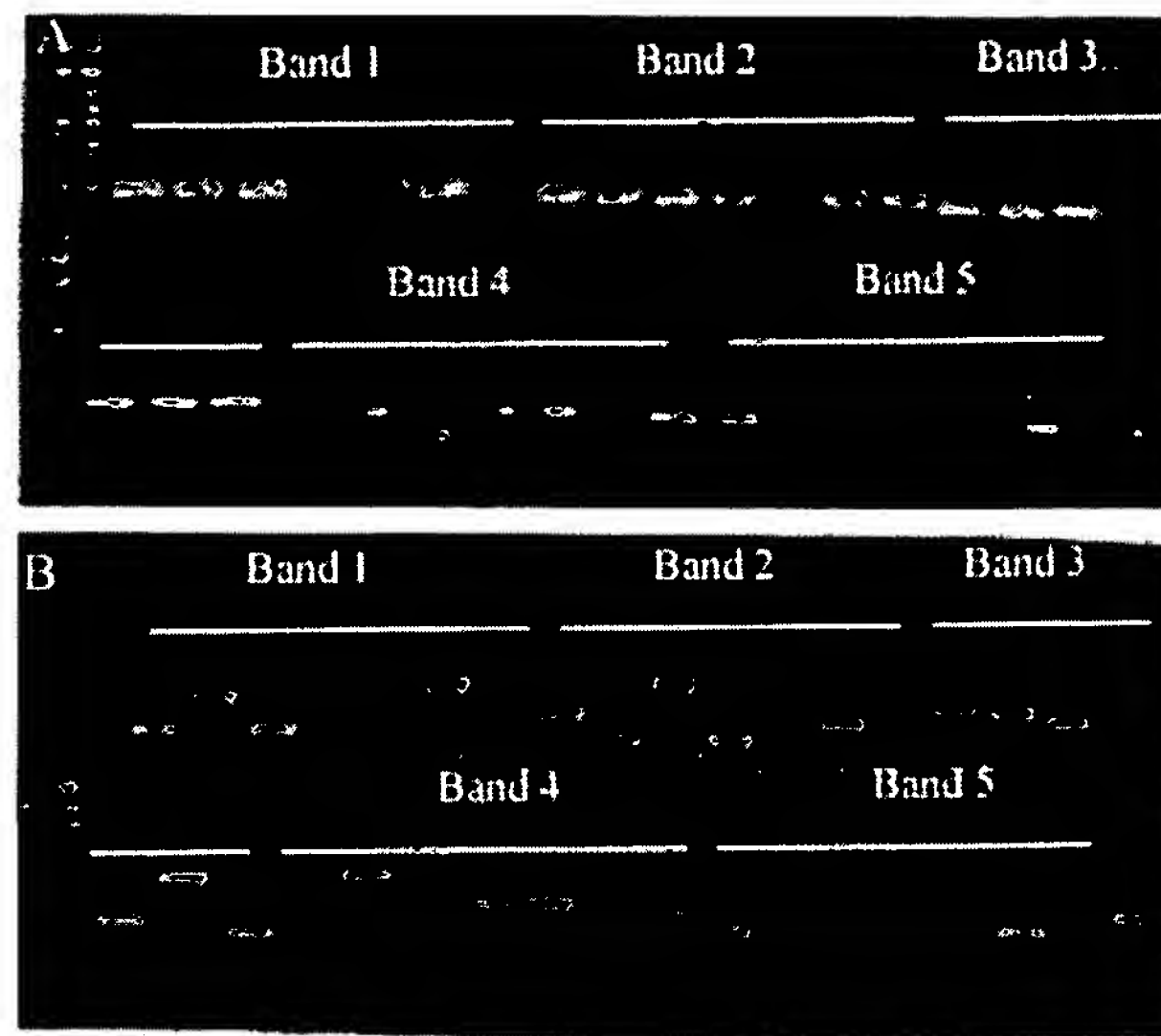


Figure 10. Discrimination of clones of identical/nearly identical size using HA-red. Bands of decreasing size (1–5) were extracted from the final display of a suppression subtractive hybridization experiment and cloned. Seven colonies were picked at random from each cloned band and their inserts amplified using PCR. The products were run on two gels, (A) a high resolution 2% agarose gel, and (B) a high resolution 2% agarose gel containing 1 U/ml HA-red. With few exceptions, all the clones from each band appear to be the same size (gel A). However, the presence of HA-red (gel B), which separates identically-sized DNA fragments based on the percentage of GC within the sequence, clearly indicates the presence of different gene species within each band. For example, even though all five re-amplified clones of band 1 appear to be the same size, at least four different gene species are represented.

in a similar gel containing one of the HA-stains. The standard gel should indicate any gross size differences, whilst the HA-stained gel should separate otherwise unresolvable species (on standard AGE) according to their base content. Geisinger *et al.* (1997) reported successful use of this approach for identifying DD-derived clones. Figure 10 shows such an experiment carried out in this laboratory on clones obtained from a band extracted from an SSH display.

An alternative approach is to carry out a 2-D analysis of the differential display products. In this approach, size-based separation is first carried out in a standard agarose gel. The gel slice containing the display is then extracted and incorporated in to a HA gel for resolution based on AT/GC content.

Of course, one should always consider the possibility of there being different gene species which are the same size and have the same GC/AT content. However, even these species are not unresolvable given some effort—again, one might use SSCP, or perhaps a denaturing gradient gel electrophoresis (DGGE) or temperature gradient field electrophoresis (TGGE) approach to resolve the contents of a band, either directly on the extracted band (Suzuki *et al.* 1991) or on the reamplified product.

The requirement of some differential display techniques to visualize large numbers of products (e.g. DD and GEF) can also present a problem in that, in terms of numbers, the resolution of PAGE rarely exceeds 300–400 bands. One approach to overcoming this might be to use 2-D gels such as those described by Uitterlinden *et al.* (1989) and Hatada *et al.* (1991).

Extraction of differentially expressed bands from a gel can be complex since, in some cases (e.g. DD, GEF), the results are visualized by autoradiographic means, such that precise overlay of the developed film on the gel must occur if the correct band is to be extracted for further analysis. Clearly, a misjudged extraction can account for many man-hours lost. This problem, and that of the use of radioisotopes, has been addressed by several groups. For example, Lohmann *et al.* (1995) demonstrated that silver staining can be used directly to visualize DD bands in horizontal PAGs. An *et al.* (1996) avoided the use of radioisotopes by transferring a small amount (20–30%) of the DNA from their DD to a nylon membrane, and visualizing the bands using chemiluminescent staining before going back to extract the remaining DNA from the gel. Chen and Peck (1996) went one step further and transferred the entire DD to a nylon membrane. The DNA bands were then visualized using a digoxigenin (DIG) system (DIG was attached to the polydT primers used in the differential display procedure). Differentially expressed bands were cut from the membrane and the DNA eluted by washing with PCR buffer prior to reamplification.

One of the advantages of using techniques such as SSH and RDA is that the final display can be run on an agarose gel and the bands visualized with simple ethidium bromide staining. Whilst this approach can provide acceptable results, overstaining with SYBR Green I or SYBR Gold nucleic acid stains (FMC) effectively enhances the intensity and sharpness of the bands. This greatly aids in their precise extraction and often reveals some faint products that may otherwise be overlooked. Whilst differential displays stained with SYBR Green I are better visualized using short wavelength UV (254 nm) rather than medium wavelength (306 nm), the shorter wavelength is much more DNA damaging. In practice, it takes only a few seconds to damage DNA extracted under 254 nm irradiation, effectively preventing reamplification and cloning. The best approach is to over stain with SYBR Green I and extract bands under a medium wavelength UV transillumination.

The possible use of 'microfingerprinting' to reduce complexity

Given the sheer number of gene products and the possible complexity of each band, an alternative approach to rapid characterization may be to use an enhanced analysis of a small section of a differential display—a 'sub-fingerprint' or 'micro-fingerprint'. In this case, one could concentrate on those bands which only appear in a particular chosen size region. Reducing the fingerprint in this way has at least two advantages. One is that it should be possible to use different gel types, concentrations and run times tailored exactly to that region. Currently, one might run products from 100–3000 + bp on the same gel, which leads to compromise in the gel system being used and consequently to suboptimal resolution, both in terms of size and numbers, and can lead to problems in the accurate excision of individual bands. Secondly, it may be possible to enhance resolution by using a 2-D analysis using a HA-stain, as described earlier. In summary, if a range of gene product sizes is carefully chosen to include certain 'relevant' genes, the 2-D system standardized, and appropriate gene analysis used, it may be possible to develop a method for the early and rapid identification of compounds which have similar or widely different cellular effects. If the prognosis for exposure to one or more other chemicals which display a similar profile is already known, then one could perhaps predict similar effects for any new compounds which show a similar micro-fingerprint.

An alternative approach to microfingerprinting is to examine altered expression in specific families of genes through careful selection of PCR primers and/or post-reaction analysis. Stress genes, growth factors and/or their receptors, cell cycling genes, cytochromes P450 and regulatory proteins might be considered as candidates for analysis in this way. Indeed, some off-the-shelf DNA arrays (e.g. Clontech's Atlas cDNA Expression Array series) already anticipated this to some degree by grouping together genes involved in different responses e.g. apoptosis, stress, DNA-damage response etc.

Screening

False positives

The generation of false positives has been discussed at length amongst the differential display community (Liang *et al.* 1993, 1995, Nishio *et al.* 1994, Sun *et al.* 1994, Sompayrac *et al.* 1995). The reason for false positives varies with the technique being used. For instance, in RDA, the use of adaptors which have not been HPLC purified can lead to the production of false positives through illegitimate ligation events (O'Neill and Sinclair 1997), whilst in DD they can arise through PCR artifacts and illegitimate transcription of rRNA. In SH, false positives appear to be derived largely from abundant gene species, although some may arise from cDNA/mRNA species which do not undergo hybridization for technical reasons.

A quick screening of putative differentially expressed clones can be carried out using a simple dot blot approach, in which labelled first strand probes synthesized from tester and driver mRNA are hybridized to an array of said clones (Hedrick *et al.* 1984, Sakaguchi *et al.* 1986). Differentially expressed clones will hybridize to tester probe, but not driver. The disadvantage of this approach is that rare species may not generate detectable hybridization signals. One option for those using SSH is to screen the clones using a labelled probe generated from the subtracted cDNA from which it was derived, and with a probe made from the reverse subtraction reaction (ClonTechniques 1997a). Since the SSH method enriches rare sequences, it should be possible to confirm the presence of clones representing low abundance genes. Despite this quick screening step, there is still the need to go back to the original mRNA and confirm the altered expression using a more quantitative approach. Although this may be achieved using Northern blots, the sensitivity is poor by today's high standards and one must rely on PCR methods for accurate and sensitive determinations (see below).

Sequence analysis

The majority of differential display procedures produce final products which are between 100 and 1000bp in size. However, this may considerably reduce the size of the sequence for analysis of the DNA databases. This in turn leads to a reduced confidence in the result—several families of genes have members whose DNA sequences are almost identical except in a few key stretches, e.g. the cytochrome P450 gene superfamily (Nelson *et al.* 1996). Thus, does the clone identified as being almost identical to gene X_0 really come from that gene, or its brother gene X_1 or its as yet undiscovered sister X_2 ? For example, using SSH, part of a gene was isolated,

which was up-regulated in the liver of rats exposed to Wy-14,643 and was identified by a FASTA search as being transferrin (data not shown). However, transferrin is known to be downregulated by hypolipidemic peroxisome proliferators such as Wy-14,643 (Hertz *et al.* 1996), and this was confirmed with subsequent RT-PCR analysis. This suggests that the gene sequence isolated may belong to a gene which is closely related to transferrin, but is regulated by a different mechanism.

A further problem associated with SH technology is redundancy. In most cases before SH is carried out, the cDNA population must first be simplified by restriction digestion. This is important for at least two reasons:

- (1) To reduce complexity—long cDNA fragments may form complex networks which prevent the formation of appropriate hybrids, especially at the high concentrations required for efficient hybridization.
- (2) Cutting the cDNAs into small fragments provides better representation of individual genes. This is because genes derived from related but distinct members of gene families often have similar coding sequences that may cross-hybridize and be eliminated during the subtraction procedure (Ko 1990). Furthermore, different fragments from the same cDNA may differ considerably in terms of hybridization and amplification and, thus, may not efficiently do one or the other (Wang and Brown 1991). Thus, some fragments from differentially expressed cDNAs may be eliminated during subtractive hybridization procedures. However, other fragments may be enriched and isolated. As a consequence of this, some genes will be cut one or more times, giving rise to two or more fragments of different sizes. If those same genes are differentially expressed, then two or more of the different size fragments may come through as separate bands on the final differential display, increasing the observed redundancy and increasing the number of redundant sequencing reactions.

Sequence comparisons also throw up another important point—at what degree of sequence similarity does one accept a result. Is 90% identity between a gene derived from your model species and another acceptably close? Is 95% between your sequence and one from the same species also acceptable? This problem is particularly relevant when the forward and reverse sequence comparisons give similar sequences with completely different gene species! An arbitrary decision seems to be to allocate genes that are definite (95% and above similarity) and then group those between 60 and 95% as being related or possible homologues.

Quantitative analysis

At some point, one must give consideration to the quantitative analysis of the candidate genes, either as a means of confirming that they are truly differentially expressed, or in order to establish just what the differences are. Northern blot analysis is a popular approach as it is relatively easy and quick to perform. However, the major drawback with Northern blots is that they are often not sensitive enough to detect rare sequences. Since the majority of messages expressed in a cell are of low abundance (see table 1), this is a major problem. Consequently, RT-PCR may be the method of choice for confirming differential expression. Although the procedure is somewhat more complex than Northern analysis, requiring synthesis of primers and optimization of reaction conditions for each gene species, it is now possible to set up high throughput PCR systems using multichannel pipettes, 96 +well plates and

appropriate thermal cycling technology. Whilst quantitative analysis is more desirable, being more accurate and without reliance on an internal standard, the money and time needed to develop a competitor molecule is often excessive, especially when one might be examining tens or even hundreds of gene species. The use of semi-quantitative analysis is simpler, although still relatively involved. One must first of all choose an internal standard that does not change in the test cells compared to the controls. Numerous reference genes have been tried in the past, for example interferon-gamma (IFN- γ , Frye *et al.* 1989), β -actin (Heuval *et al.* 1994), glyceraldehyde-3-phosphate dehydrogenase (GAPDH, Wong *et al.* 1994), dihydrofolate reductase (DHFR, Mohler and Butler 1991), β -2-microglobulin (β -2-m, Murphy *et al.* 1990), hypoxanthine phosphoribosyl transferase (HPRT, Foss *et al.* 1998) and a number of others (ClonTechniques 1997b). Ideally, an internal standard should not change its level of expression in the cell regardless of cell age, stage in the cell cycle or through the effects of external stimuli. However, it has been shown on numerous occasions that the levels of most housekeeping genes currently used by the research community do in fact change under certain conditions and in different tissues (ClonTechniques 1997b). It is imperative, therefore, that preliminary experiments be carried out on a panel of housekeeping genes to establish their suitability for use in the model system.

Interpretation of quantitative data must also be treated with caution. By comparing the lists of genes identified by differential expression one can perhaps gain insight into why two different species react in different ways to external stimuli. For example, rats and mice appear sensitive to the non-genotoxic effects of a wide range of peroxisome proliferators whilst Syrian hamsters and guinea pigs are largely resistant (Orton *et al.* 1984, Rodricks and Turnbull 1987, Lake *et al.* 1989, 1993, Makowska *et al.* 1992). A simplified approach to resolving the reason(s) why is to compare lists of up- and down-regulated genes in order to identify those which are expressed in only one species and, through background knowledge of the effects of the said gene, might suggest a mechanism of facilitated non-genotoxic carcinogenesis or protection. Of course, the situation is likely to be far more complex. Perhaps if there were one key gene protecting guinea pig from non-genotoxic effects and it was upregulated 50 times by PPs, the same gene might only be up-regulated five times in the rat. However, since both were noted to be upregulated, the importance of the gene may be overlooked. Just to complicate matters, a large change in expression does not necessarily mean a biologically important change. For example, what is the true relevance of gene Y which shows a 50-fold increase after a particular treatment, and gene Z which shows only a 5-fold increase? If one examines the literature one may find that historically, gene Y has often been shown to be up-regulated 40–60-fold by a number of unrelated stimuli—in light of this the 50-fold increase would appear less significant. However, the literature may show that gene Z has never been recorded as having more than doubled in expression—which makes your 5-fold increase all the more exciting. Perhaps even more interesting is if that same 5-fold increase has only been seen in related neoplasms or following treatment with related chemicals.

Problems in using the differential display approach

Differential display technology originally held promise of an easily obtainable 'fingerprint' of those genes which are up- or down-regulated in test animals/cells in a developmental process or following exposure to given stimuli. However, it has

become clear that the fingerprinting process, whilst still valid, is much too complex to be represented by a single technique profile. This is because all differential display techniques have common and/or unique technical problems which preclude the isolation and identification of all those genes which show changes in expression. Furthermore, there are important genetic changes related to disease development which differential expression analysis is simply not designed to address. An example of this is the presence of small deletions, insertions, or point mutations such as those seen in activated oncogenes, tumour suppressor genes and individual polymorphisms. Polymorphic variations, small though they usually are, are often regarded as being of paramount importance in explaining why some patients respond better than others to certain drug treatments (and, in logical extension, why some people are less affected by potentially dangerous xenobiotics/carcinogens than others). The identification of such point mutations and naturally occurring polymorphisms requires the subsequent application of sequencing, SSCP, DGGE or TGGE to the gene of interest. Furthermore, differential display is not designed to address issues such as alternatively spliced gene species or whether an increased abundance of mRNA is a result of increased transcription or increased mRNA stability.

Conclusions

Perhaps the main advantage of open system differential display techniques is that they are not limited by extant theories or researcher bias in revealing genes which are differentially expressed, since they are designed to amplify all genes which demonstrate altered expression. This means that they are useful for the isolation of previously unknown genes which may turn out be useful biomarkers of a particular state or condition. At least one open system (SAGE) is also quantitative, thus eliminating the need to return to the original mRNA and carry out Northern/PCR analysis to confirm the result. However, the rapid progress of genome mapping projects means that over the next 5–10 years or so, the balance of experimental use will switch from open to closed differential display systems, particularly DNA arrays. Arrays are easier and faster to prepare and use, provide quantitative data, are suitable for high throughput analysis and can be tailored to look at specific signalling pathways or families of genes. Identification of all the gene sequences in human and common laboratory animals combined with improved DNA array technology, means that it will soon no longer be necessary to try to isolate differentially expressed genes using the technically more demanding open system approach. Thus, their main advantage (that of identifying unknown genes) will be largely eradicated. It is likely, therefore, that their sphere of application will be reduced to analysis of the less common laboratory species, since it will be some time yet before the genomes of such animals as zebrafish, electric eels, gerbils, crayfish and squid, for example, will be sequenced.

Of course, in the end the question will always remain: What is the functional/biological significance of the identified, differentially expressed genes? One persistent problem is understanding whether differentially expressed genes are a cause or consequence of the altered state. Furthermore, many chemicals, such as non-genotoxic carcinogens, are also mitogens and so genes associated with replication will also be upregulated but may have little or nothing to do with the

carcinogenic effect. Whilst differential display technology cannot hope to answer these questions, it does provide a springboard from which identification, regulatory and functional studies can be launched. Understanding the molecular mechanism of cellular responses is almost impossible without knowing the regulation and function of those genes and their condition (e.g. mutated). In an abstract sense, differential display can be likened to a still photograph, showing details of a fixed moment in time. Consider the Historian who knows the outcome of a battle and the placement and condition of the troops before the battle commenced, but is asked to try and deduce how the battle progressed and why it ended as it did from a few still photographs—an impossible task. In order to understand the battle, the Historian must find out the capabilities and motivation of the soldiers and their commanding officers, what the orders were and whether they were obeyed. He must examine the terrain, the remains of the battle and consider the effects the prevailing weather conditions exerted. Likewise, if mechanistic answers are to be forthcoming, the scientist must use differential display in combination with other techniques, such as knockout technology, the analysis of cell signalling pathways, mutation analysis and time and dose response analyses. Although this review has emphasized the importance of differential gene profiling, it should not be considered in isolation and the full impact of this approach will be strengthened if used in combination with functional genomics and proteomics (2-dimensional protein gels from isoelectric focusing and subsequent SDS electrophoresis and virtual 2D-maps using capillary electrophoresis). Proteomics is attracting much recent attention as many of the changes resulting in differential gene expression do not involve changes in mRNA levels, as described extensively herein, but rather protein–protein, protein–DNA and protein phosphorylation events which would require functional genomics or proteomic technologies for investigation.

Despite the limitations of differential display technology, it is clear that many potential applications and benefits can be obtained from characterizing the genetic changes that occur in a cell during normal and disease development and in response to chemical or biological insult. In light of functional data, such profiling will provide a 'fingerprint' of each stage of development or response, and in the long term should help in the elucidation of specific and sensitive biomarkers for different types of chemical/biological exposure and disease states. The potential medical and therapeutic benefits of understanding such molecular changes are almost immeasurable. Amongst other things, such fingerprints could indicate the family or even specific type of chemical an individual has been exposed to plus the length and/or acuteness of that exposure, thus indicating the most prudent treatment. They may also help uncover differences in histologically identical cancers, provide diagnostic tests for the earliest stages of neoplasia and, again, perhaps indicate the most efficacious treatment.

The Human Genome Project will be completed early in the next century and the DNA sequence of all the human genes will be known. The continuing development and evolution of differential gene expression technology will ensure that this knowledge contributes fully to the understanding of human disease processes.

Acknowledgements

We acknowledge Drs Nick Plant (University of Surrey), Sally Darney and Chris Luft (US EPA at RTP) for their critical analysis of the manuscript prior to submission. This manuscript has been reviewed in accordance with the policy of the

US Environmental Protection Agency and approved for publication. Approval does not signify that the contents reflect the views and policies of the Agency, nor does mention of trade names constitute endorsement or recommendation for use.

References

- ADAMS, M. D., KELLEY, J. M., GOCAYNE, J. D., DUBNICK, M., POLYMERPOULOS, M. H., XIAO, H., MERRIL, C. R., WU, A., OLDE, B., MORENO, R. F., KERLAVAGE, A. R., MCCOMBIE, W. R. and VENTOR, J. C., 1991, Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651–1656.
- AN, G., LUO, G., VELTRI, R. W. and O'HARA, S. M., 1996, Sensitive non-radioactive differential display method using chemiluminescent detection. *Biotechniques*, **20**, 342–346.
- AXEL, R., FEIGELSON, P. and SCHULTZ, G., 1976, Analysis of the complexity and diversity of mRNA from chicken liver and oviduct. *Cell*, **7**, 247–254.
- BAND, V. and SAGER, R., 1989, Distinctive traits of normal and tumor-derived human mammary epithelial cells expressed in a medium that supports long-term growth of both cell types. *Proceedings of the National Academy of Sciences, USA*, **86**, 1249–1253.
- BAUER, D., MULLER, H., REICH, J., RIEDEL, H., AHRENKIEL, V., WARTHOF, P. and STRAUSS, M., 1993, Identification of differentially expressed mRNA species by an improved display technique (DDRT-PCR). *Nucleic Acids Research*, **21**, 4272–4280.
- BERTIOLI, D. J., SCHLICHTER, U. H. A., ADAMS, M. J., BURROWS, P. R., STEINBISS, H.-H. and ANTONIW, J. F., 1995, An analysis of differential display shows a strong bias towards high copy number mRNAs. *Nucleic Acids Research*, **23**, 4520–4523.
- BRAVO, R., 1990, Genes induced during the G0/G1 transition in mouse fibroblasts. *Seminars in Cancer Biology*, **1**, 37–46.
- BURN, T. C., PETROVICK, M. S., HOHAUS, S., ROLLINS, B. J. and TENEN, D. G., 1994, Monocyte chemoattractant protein-1 gene is expressed in activated neutrophils and retinoic acid-induced human myeloid cell lines. *Blood*, **84**, 2776–2783.
- CAO, J., CAI, X., ZHENG, L., GENG, L., SHI, Z., PAO, C. C. and ZHENG, S., 1997, Characterisation of colorectal cancer-related cDNA clones obtained by subtractive hybridisation screening. *Journal of Cancer Research and Clinical Oncology*, **123**, 447–451.
- CASSIDY, S. B., 1995, Uniparental disomy and genomic imprinting as causes of human genetic disease. *Environmental and Molecular Mutagenesis*, **25** (Suppl 26), 13–20.
- CHANG, G. W. and TERZAGHI-HOWE, M., 1998, Multiple changes in gene expression are associated with normal cell-induced modulation of the neoplastic phenotype. *Cancer Research*, **58**, 4445–4452.
- CHEN, J., SCHWARTZ, D. A., YOUNG, T. A., NORRIS, J. S. and YAGER, J. D., 1996, Identification of genes whose expression is altered during mitosuppression in livers of ethinyl estradiol-treated female rats. *Carcinogenesis*, **17**, 2783–2786.
- CHEN, J. J. W. and PECK, K., 1996, Non-radioactive differential display method to directly visualise and amplify differential bands on nylon membrane. *Nucleic Acid Research*, **24**, 793–794.
- CLON TECHNIQUES, 1997a, PCR-Select Differential Screening Kit—the nextstep after Clontech PCR-Select cDNA subtraction. *ClonTechniques*, **XII**, 18–19.
- CLON TECHNIQUES, 1997b, Housekeeping RT-PCR amplimers and cDNA probes. *ClonTechniques*, **XII**, 15–16.
- DAVIS, M. M., COHEN, D. I., NIELSEN, E. A., STEINMETZ, M., PAUL, W. E. and HOOD, L., 1984, Cell-type-specific cDNA probes and the murine I region: the localization and orientation of Ad alpha. *Proceedings of the National Academy of Sciences (USA)*, **81**, 2194–2198.
- DELLAVALLE, R. P., PETERSON, R. and LINDQUIST, S., 1994, Preferential deadenylation of HSP70 mRNA plays a key role in regulating Hsp70 expression in *Drosophila melanogaster*. *Molecular and Cell Biology*, **14**, 3646–3659.
- DERISI, J. L., VASHWANATH, R. L. and BROWN, P., 1997, Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.
- DIATCHENKO, L., LAU, Y.-F. C., CAMPBELL, A. P., CHENCHIK, A., MOQADAM, F., HUANG, B., LUKYANOV, K., GURSKAYA, N., SVERDLOV, E. D. and SIEBERT, P. D., 1996, Suppression subtractive hybridisation: A method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proceedings of the National Academy of Sciences (USA)*, **93**, 6025–6030.
- DOGRA, S. C., WHITELAW, M. L. and MAY, B. K., 1998, Transcriptional activation of cytochrome P450 genes by different classes of chemical inducers. *Clinical and Experimental Pharmacology and Physiology*, **25**, 1–9.
- DUGUID, J. R. and DINAUER, M. C., 1990, Library subtraction of *in vitro* cDNA libraries to identify differentially expressed genes in scrapie infection. *Nucleic Acids Research*, **18**, 2789–2792.
- DUNBAR, P. R., OGG, G. S., CHEN, J., RUST, N., VAN DER BRUGGEN, P. and CERUNDOLO, V., 1998, Direct isolation, phenotyping and cloning of low-frequency antigen-specific cytotoxic T lymphocytes from peripheral blood. *Current Biology*, **26**, 413–416.

- FITZPATRICK, D. R., GERMAIN -LEE, E. and VALLE, D., 1995, Isolation and characterisation of rat and human cDNAs encoding a novel putative peroxisomal enoyl-CoA hydratase. *Genomics*, **27**, 457-466.
- FOSS, D. L., BAARSCH, M. J. and MURTAUGH, M. P., 1998, Regulation of hypoxanthine phosphoribosyltransferase, glyceraldehyde-3-phosphate dehydrogenase and beta-actin mRNA expression in porcine immune cells and tissues. *Animal Biotechnology*, **9**, 67-78.
- FRYE, R. A., BENZ, C. C. and LIU, E., 1989, Detection of amplified oncogenes by differential polymerase chain reaction. *Oncogene*, **4**, 1153-1157.
- GEISINGER, A., RODRIGUEZ, R., ROMERO, V. and WETTSTEIN, R., 1997, A simple method for screening cDNAs arising from the cloning of RNA differential display bands. *Elsevier Trends Journals Technical Tips Online*, <http://tto.trends.com>, document T01110.
- GRESS, T. M., HOHEISEL, J. D., LENNON, G. G., ZEHETNER, G. and LEHRACH, H., 1992, Hybridisation fingerprinting of high density cDNA filter arrays with cDNA pools derived from whole tissues. *Mammalian Genome*, **3**, 609-619.
- GRIFFIN, G. and KRISHNA, S., 1998, Cytokines in infectious diseases. *Journal of the Royal College of Physicians, London*, **32**, 195-198.
- GROENINK, M. and LEEGWATER, A. C. J., 1996, Isolation of delayed early genes associated with liver regeneration using Clontech PCR-select subtraction technique. *Clontechniques*, **XI**, 23-24.
- GUIMARAES, M. J., BAZAN, J. F., ZLOTNIK, A., WILES, M. V., GRIMALDI, J. C., LEE, F. and MCCLANAHAN, T., 1995b, A new approach to the study of haematopoietic development in the yolk sac and embryoid bodies. *Development*, **121**, 3335-3346.
- GUIMARAES, M. J., LEE, F., ZLOTNIK, A. and MCCLANAHAN, T., 1995a, Differential display by PCR: novel findings and applications. *Nucleic Acids Research*, **23**, 1832-1833.
- GURSKAYA, N. G., DIATCHENKO, L., CHENCHIK, P. D., SIEBERT, P. D., KHASPEKOV, G. L., LUKYANOV, K. A., VAGNER, L. L., ERMOLAEVA, O. D., LUKYANOV, S. A. and SVERDLOV, E. D., 1996, Equalising cDNA subtraction based on selective suppression of polymerase chain reaction: Cloning of Jurkat cell transcripts induced by phytohemagglutinin and phorbol 12-Myristate 13-Acetate. *Analytical Biochemistry*, **240**, 90-97.
- HAMPSON, I. N. and HAMPSON, L., 1997, CCLS and DROP—subtractive cloning made easy. *Life Science News* (A publication of Amersham Life Science), **23**, 22-24.
- HAMPSON, I. N., HAMPSON, L. and DEXTER, T. M., 1996, Directional random oligonucleotide primed (DROP) global amplification of cDNA: its application to subtractive cDNA cloning. *Nucleic Acids Research*, **24**, 4832-4835.
- HAMPSON, I. N., POPE, L., COWLING, G. J. and DEXTER, T. M., 1992, Chemical cross linking subtraction (CCLS): a new method for the generation of subtractive hybridisation probes. *Nucleic Acids Research*, **20**, 2899.
- HARA, E., KATO, T., NAKADA, S., SEKIYA, S. and ODA, K., 1991, Subtractive cDNA cloning using oligo(dT)30-latex and PCR: isolation of cDNA clones specific to undifferentiated human embryonal carcinoma cells. *Nucleic Acids Research*, **19**, 7097-7104.
- HATADA, I., HAYASHIZAKE, Y., HIROTSUNE, S., KOMATSUBARA, H. and MUKAI, T., 1991, A genomic scanning method for higher organisms using restriction sites as landmarks. *Proceedings of the National Academy of Sciences (USA)*, **88**, 9523-9527.
- HECHT, N., 1998, Molecular mechanisms of male sperm cell differentiation. *Bioessays*, **20**, 555-561.
- HEDRICK, S., COHEN, D. I., NIELSEN, E. A. and DAVIS, M. E., 1984, Isolation of T cell-specific membrane-associated proteins. *Nature*, **308**, 149-153.
- HERTZ, R., SECKBACH, M., ZAKIN, M. M. and BAR-TANA, J., 1996, Transcriptional suppression of the transferrin gene by hypolipidemic peroxisome proliferators. *Journal of Biological Chemistry*, **271**, 218-224.
- HEUVAL, J. P. V., CLARK, G. C., KOHN, M. C., TRITSCHER, A. M., GREENLEE, W. F., LUCIER, G. W. and BELL, D. A., 1994, Dioxin-responsive genes: Examination of dose-response relationships using quantitative reverse transcriptase-polymerase chain reaction. *Cancer Research*, **54**, 62-68.
- HILLIER, L. D., LENNON, G., BECKER, M., BONALDO, M. F., CHIAPELLI, B., CHISSOE, S., DIETRICH, N., DUBUQUE, T., FAVELLO, A., GISH, W., HAWKINS, M., HULTMAN, M., KUCABA, T., LACY, M., LE, M., LE, N., MARDIS, E., MOORE, B., MORRIS, M., PARSONS, J., PRANGE, C., RIFKIN, L., ROHLFING, T., SCHELLENBERG, K., SOARES, M. B., TAN, F., THIERRY-MEG, J., TREVASKIS, E., UNDERWOOD, K., WOHLDMAN, P., WATERSTON, R., WILSON, R. and MARRA, M., 1996, Generation and analysis of 280,000 human expressed sequence tags. *Genome Research*, **6**, 807-828.
- HUBANK, M. and SCHATZ, D. G., 1994, Identifying differences in mRNA expression by representational difference analysis. *Nucleic Acids Research*, **22**, 5640-5648.
- HUNTER, T., 1991, Cooperation between oncogenes. *Cell*, **64**, 249-270.
- IVANOVA, N. B. and BELYAVSKY, A. V., 1995, Identification of differentially expressed genes by restriction endonuclease-based gene expression fingerprinting. *Nucleic Acids Research*, **23**, 2954-2958.
- JAMES, B. D. and HIGGINS, S. J., 1985, *Nucleic Acid Hybridisation* (Oxford: IRL Press Ltd).
- KAS-DEELEN, A. M., HARMSSEN, M. C., DE MAAR, E. F. and VAN SON, W. J., 1998, A sensitive method for

- quantifying cytomegalic endothelial cells in peripheral blood from cytomegalovirus-infected patients. *Clinical Diagnostic and Laboratory Immunology*, 5, 622-626.
- KILTY, I. and VICKERS, P., 1997, Fractionating DNA fragments generated by differential display PCR. *Strategies Newsletter* (Stratagene), 10, 50-51.
- KLEINJAN, D.-J. and VAN HEYNINGEN, V., 1998, Position effect in human genetic disease. *Human and Molecular Genetics*, 7, 1611-1618.
- KO, M. S., 1990, An 'equalized cDNA library' by the reassociation of short double-stranded cDNAs. *Nucleic Acids Research*, 18, 5705-5711.
- LAKE, B. G., EVANS, J. G., CUNNINGHAME, M. E. and PRICE, R. J., 1993, Comparison of the hepatic effects of Wy-14,643 on peroxisome proliferation and cell replication in the rat and Syrian hamster. *Environmental Health Perspectives*, 101, 241-248.
- LAKE, B. G., EVANS, J. G., GRAY, T. J. B., KOROSI, S. A. and NORTH, C. J., 1989, Comparative studies of nafenopin-induced hepatic peroxisome proliferation in the rat, Syrian hamster, guinea pig and marmoset. *Toxicology and Applied Pharmacology*, 99, 148-160.
- LENNARD, M. S., 1993, Genetically determined adverse drug reactions involving metabolism. *Drug Safety*, 9, 60-77.
- LEVY, S., TODD, S. C. and MAECKER, H. T., 1998, CD81(TAPA-1): a molecule involved in signal transduction and cell adhesion in the immune system. *Annual Review of Immunology*, 16, 89-109.
- LIANG, P. and PARDEE, A. B., 1992, Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, 257, 967-971.
- LIANG, P., AVERBOUKH, L., KEYOMARSI, K., SAGER, R. and PARDEE, A., 1992, Differential display and cloning of messenger RNAs from human breast cancer versus mammary epithelial cells. *Cancer Research*, 52, 6966-6968.
- LIANG, P., AVERBOUKH, L. and PARDEE, A. B., 1993, Distribution & cloning of eukaryotic mRNAs by means of differential display refinements and optimisation. *Nucleic Acids Research*, 21, 3269-3275.
- LIANG, P., BAUER, D., AVERBOUKH, L., WARTHOF, P., ROHRWILD, M., MULLER, H., STRAUSS, M. and PARDEE, A. B., 1995, Analysis of altered gene expression by differential display. *Methods in Enzymology*, 254, 304-321.
- LINSKENS, M. H., FENG, J., ANDREWS, W. H., ENLOW, B. E., SAATI, S. M., TONKIN, L. A., FUNK, W. D. and VILLEPONTEAU, B., 1995, Cataloging altered gene expression in young and senescent cells using enhanced differential display. *Nucleic Acids Research*, 23, 3244-3251.
- LISITSYN, N., LISITSYN, N. and WIGLER, M., 1993, Cloning the differences between two complex genomes. *Science*, 259, 946-951.
- LOHMANN, J., SCHICKLE, H. and BOSCH, T. C. G., 1995, REN Display, a rapid and efficient method for non-radioactive differential display and mRNA isolation. *Biotechniques*, 18, 200-202.
- LUNNEY, J. K., 1998, Cytokines orchestrating the immune response. *Reviews in Science and Technology*, 17, 84-94.
- MAKOWSKA, J. M., GIBSON, G. G. and BONNER, F. W., 1992, Species differences in ciprofibrate-induction of hepatic cytochrome P450A1 and peroxisome proliferation. *Journal of Biochemical Toxicology*, 7, 183-191.
- MALDARELLI, F., XIANG, C., CHAMOUN, G. and ZEICHNER, S. L., 1998, The expression of the essential nuclear splicing factor SC35 is altered by human immunodeficiency virus infection. *Virus Research*, 53, 39-51.
- MATHIEU-DAUDE, F., CHENG, R., WELSH, J. and MCCLELLAND, M., 1996, Screening of differentially amplified cDNA products from RNA arbitrarily primed PCR fingerprints using single strand conformation polymorphism (SSCP) gels. *Nucleic Acids Research*, 24, 1504-1507.
- MCKENZIE, D. and DRAKE, D., 1997, Identification of differentially expressed gene products with the castaway system. *Strategies Newsletter* (Stratagene), 10, 19-20.
- MCCLELLAND, M., MATHIEU-DAUDE, F. and WELSH, J., 1996, RNA fingerprinting and differential display using arbitrarily primed PCR. *Trends in Genetics*, 11, 242-246.
- MECHLER, B. and RABBITTS, T. H., 1981, Membrane-bound ribosomes of myeloma cells. IV. mRNA complexity of free and membrane-bound polysomes. *Journal of Cell Biology*, 88, 29-36.
- MEYER, U. A. and ZANGER, U. M., 1997, Molecular mechanisms of genetic polymorphisms of drug metabolism. *Annual Review of Pharmacology and Toxicology*, 37, 269-296.
- MOHLER, K. M. and BUTLER, L. D., 1991, Quantitation of cytokine mRNA levels utilizing the reverse transcriptase-polymerase chain reaction following primary antigen-specific sensitization in vivo—I. Verification of linearity, reproducibility and specificity. *Molecular Immunology*, 28, 437-447.
- MURPHY, L. D., HERZOG, C. E., RUDICK, J. B., TITO FOJO, A. and BATES, S. E., 1990, Use of the polymerase chain reaction in the quantitation of the *mdr-1* gene expression. *Biochemistry*, 29, 10351-10356.
- NELSON, D. R., KOYMANS, L., KAMATAKI, T., STEGEMAN, J. J., FEYEREISEN, R., WAXMAN, D. J., WATERMAN, M. R., GOTOH, O., COON, M. J., ESTABROOK, R. W., GUNSALUS, I. C. and NEBERT, D. W., 1996, Update on new sequences, gene mapping, accession numbers and nomenclature. *Pharmacogenetics*, 6, 1-42.

- NISHIO, Y., AIELLO, L. P. and KING, G. L., 1994, Glucose induced genes in bovine aortic smooth muscle cells identified by mRNA differential display. *FASEB Journal*, **8**, 103–106.
- O'NEILL, M. J. and SINCLAIR, A. H., 1997, Isolation of rare transcripts by representational difference analysis. *Nucleic Acids Research*, **25**, 2681–2682.
- ORTON, T. C., ADAM, H. K., BENTLEY, M., HOLLOWAY, B. and TUCKER, M. J., 1984, Clobuzarit: species differences in the morphological and biochemical response of the liver following chronic administration. *Toxicology and Applied Pharmacology*, **73**, 138–151.
- PELKONEN, O., MAENPAA, J., TAAVITSAINEN, P., RAUTIO, A. and RAUNIO, H., 1998, Inhibition and Induction of human cytochrome P450 (CYP) enzymes. *Xenobiotica*, **28**, 1203–1253.
- PHILIPS, S. M., BENDALL, A. J. and RAMSHAW, I. A., 1990, Isolation of genes associated with high metastatic potential in rat mammary adenocarcinomas. *Journal of the National Cancer Institute*, **82**, 199–203.
- PRASHAR, Y. and WEISSMAN, S. M., 1996, Analysis of differential gene expression by display of 3' end restriction fragments of cDNAs. *Proceedings of the National Academy of Sciences (USA)*, **93**, 659–663.
- RAGNO, S., ESTRADA, I., BUTLER, R. and COLSTON, M. J., 1997, Regulation of macrophage gene expression following invasion by *Mycobacterium tuberculosis*. *Immunology Letters*, **57**, 143–146.
- RAMANA, K. V. and KOHLI, K. K., 1998, Gene regulation of cytochrome P450—an overview. *Indian Journal of Experimental Biology*, **36**, 437–446.
- RICHARD, L., VELASCO, P. and DETMAR, M., 1998, A simple immunomagnetic protocol for the selective isolation and long-term culture of human dermal microvascular endothelial cells. *Experimental Cell Research*, **240**, 1–6.
- ROCKETT, J. C., ESDAILE, D. J. and GIBSON, G. G., 1997, Molecular profiling of non-genotoxic hepatocarcinogenesis using differential display reverse transcription-polymerase chain reaction (ddRT-PCR). *European Journal of Drug Metabolism and Pharmacokinetics*, **22**, 329–333.
- RODRICKS, J. V. and TURNBULL, D., 1987, Inter-species differences in peroxisomes and peroxisome proliferation. *Toxicology and Industrial Health*, **3**, 197–212.
- ROGLER, G., HAUSMANN, M., VOGL, D., ASCHENBRENNER, E., ANDUS, T., FALK, W., ANDRESEN, R., SCHOLMERICH, J. and GROSS, V., 1998, Isolation and phenotypic characterization of colonic macrophages. *Clinical and Experimental Immunology*, **112**, 205–215.
- ROHN, W. M., LEE, Y. J. and BENVENISTE, E. N., 1996, Regulation of class II MHC expression. *Critical Reviews in Immunology*, **16**, 311–330.
- RUDIN, C. M. and THOMPSON, C. B., 1998, B-cell development and maturation. *Seminars in Oncology*, **25**, 435–446.
- SAKAGUCHI, N., BERGER, C. N. and MELCHERS, F., 1986, Isolation of a cDNA copy of an RNA species expressed in murine pre-B cells. *EMBO Journal*, **5**, 2139–2147.
- SAMBROOK, J., FRITSCH, E. F. and MANIATIS, T., 1989, Gel electrophoresis of DNA. In N. Ford, M. Nolan and M. Ferguson (eds), *Molecular Cloning—A laboratory manual*, 2nd edition (New York: Cold Spring Harbour Laboratory Press), Volume 1, pp. 6–37.
- SARGENT, T. D. and DAWID, I. B., 1983, Differential gene expression in the gastrula of *Xenopus laevis*. *Science*, **222**, 135–139.
- SCHEINA, M., SHALON, D., HELLER, R., CHAI, A., BROWN, P. O. and DAVIS, R. W., 1996, Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences (USA)*, **93**, 10614–10619.
- SCHNEIDER, C., KING, R. M. and PHILIPSON, L., 1988, Genes specifically expressed at growth arrest of mammalian cells. *Cell*, **54**, 787–793.
- SCHNEIDER-MAUNOURY, S., GILARDI-HEBENSTREIT, P. and CHARNAY, P., 1998, How to build a vertebrate hindbrain. Lessons from genetics. *C R Academy of Science III*, **321**, 819–834.
- SEMENZA, G. L., 1994, Transcriptional regulation of gene expression: mechanisms and pathophysiology. *Human Mutations*, **3**, 180–199.
- SEWALL, C. H., BELL, D. A., CLARK, G. C., TRITSCHER, A. M., TULLY, D. B., VANDEN HEUVEL, J. and LUCIER, G. W., 1995, Induced gene transcription: implications for biomarkers. *Clinical Chemistry*, **41**, 1829–1834.
- SINGH, N., AGRAWAL, S. and RASTOGI, A. K., 1997, Infectious diseases and immunity: special reference to major histocompatibility complex. *Emerging Infectious Diseases*, **3**, 41–49.
- SMITH, N. R., LI, A., ALDERSLEY, M., HIGH, A. S., MARKHAM, A. F. and ROBINSON, P. A., 1997, Rapid determination of the complexity of cDNA bands extracted from DDRT-PCR polyacrylamide gels. *Nucleic Acids Research*, **25**, 3552–3554.
- SOMPAYRAC, L., JANE, S., BURN, T. C., TENEN, D. G. and DANNA, K. J., 1995, Overcoming limitations of the mRNA differential display technique. *Nucleic Acids Research*, **23**, 4738–4739.
- ST JOHN, T. P. and DAVIS, R. W., 1979, Isolation of galactose-inducible DNA sequences from *Saccharomyces cerevisiae* by differential plaque filter hybridisation. *Cell*, **16**, 443–452.
- SUN, Y., HEGAMER, G. and COLBURN, N. H., 1994, Molecular cloning of five messenger RNAs differentially expressed in preneoplastic or neoplastic JB6 mouse epidermal cells: one is homologous to human tissue inhibitor of metalloproteinases-3. *Cancer Research*, **54**, 1139–1144.

- SUNG, Y. J. and DENMAN, R. B., 1997, Use of two reverse transcriptases eliminates false-positive results in differential display. *Biotechniques*, **23**, 462-464.
- SUTTON, G., WHITE, O., ADAMS, M. and KERLAVAGE, A., 1995, TIGR Assembler; A new tool for assembling large shotgun sequencing projects. *Genome Science and Technology*, **1**, 9-19.
- SUZUKI, Y., SEKIYA, T. and HAYASHI, K., 1991, Allele-specific polymerase chain reaction: a method for amplification and sequence determination of a single component among a mixture of sequence variants. *Analytical Biochemistry*, **192**, 82-84.
- SYED, V., GU, W. and HECHT, N. B., 1997, Sertoli cells in culture and mRNA differential display provide a sensitive early warning assay system to detect changes induced by xenobiotics. *Journal of Andrology*, **18**, 264-273.
- UITTERLINDEN, A. G., SLAGBOOM, P., KNOOK, D. L. and VIJGL, J., 1989, Two-dimensional DNA fingerprinting of human individuals. *Proceedings of the National Academy of Sciences (USA)*, **86**, 2742-2746.
- ULLMAN, K. S., NORTHROP, J. P., VERWEIJ, C. L. and CRABTREE, G. R., 1990, Transmission of signals from the T lymphocyte antigen receptor to the genes responsible for cell proliferation and immune function: the missing link. *Annual Review of Immunology*, **8**, 421-452.
- VASMATZIS, G., ESSAND, M., BRINKMANN, U., LEE, B. and PASTON, I., 1998, Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis. *Proceedings of the National Academy of Sciences (USA)*, **95**, 300-304.
- VELCULESCU, V. E., ZHANG, L., VOGELSTEIN, B. and KINZLER, K. W., 1995, Serial analysis of gene expression. *Science*, **270**, 484-487.
- VOELTZ, G. K. and STEITZ, J. A., 1998, AuuuA sequences direct mRNA deadenylation uncoupled from decay during *Xenopus* early development. *Molecular and Cell Biology*, **18**, 7537-7545.
- VOGELSTEIN, B. and KINZLER, K. W., 1993, The multistep nature of cancer. *Trends in Genetics*, **9**, 138-141.
- WALTER, J., BELFIELD, M., HAMPSON, I. and READ, C., 1997, A novel approach for generating subtractive probes for differential screening by CCLS. *Life Science News*, **21**, 13-14.
- WAN, J. S., SHARP, S. J., POIRIER, G. M.-C., WAGAMAN, P. C., CHAMBERS, J., PYATI, J., HOM, Y.-L., GALINDO, J. E., HUVAR, A., PETERSON, P. A., JACKSON, M. R. and ERLANDER, M. G., 1996, Cloning differentially expressed mRNAs. *Nature Biotechnology*, **14**, 1685-1691.
- WALTER, J., BELFIELD, M., HAMPSON, I. and READ, C., 1997, A novel approach for generating subtractive probes for differential screening by CCLS. *Life Science News*, **21**, 13-14.
- WANG, Z. and BROWN, D. D., 1991, A gene expression screen. *Proceedings of the National Academy of Sciences (USA)*, **88**, 11505-11509.
- WAWER, C., RUGGERBERG, H., MEYER, G. and MUYZER, G., 1995, A simple and rapid electrophoresis method to detect sequence variation in PCR-amplified DNA fragments. *Nucleic Acids Research*, **23**, 4928-4929.
- WELSH, J., CHADA, K., DALAL, S. S., CHENG, R., RALPH, D. and MCCLELLAND, M., 1992, Arbitrarily primed PCR fingerprinting of RNA. *Nucleic Acids Research*, **20**, 4965-4970.
- WONG, H., ANDERSON, W. D., CHENG, T. and RIABOWOL, K. T., 1994, Monitoring mRNA expression by polymerase chain reaction: the 'primer-dropping' method. *Analytical Biochemistry*, **223**, 251-258.
- WONG, K. K. and MCCLELLAND, M., 1994, Stress-inducible gene of *Salmonella typhimurium* identified by arbitrarily primed PCR of RNA. *Proceedings of the National Academy of Sciences (USA)*, **91**, 639-643.
- WYNFORD-THOMAS, D., 1991, Oncogenes and anti-oncogenes; the molecular basis of tumour behaviour. *Journal of Pathology*, **165**, 187-201.
- XHU, D., CHAN, W. L., LEUNG, B. P., HUANG, F. P., WHEELER, R., PIEDRAFITA, D., ROBINSON, J. H. and LIEW, F. Y., 1998, Selective expression of a stable cell surface molecule on type 2 but not type 1 helper T cells. *Journal of Experimental Medicine*, **187**, 787-794.
- YANG, M. and SYTOWSKI, A. J., 1996, Cloning differentially expressed genes by linker capture subtraction. *Analytical Biochemistry*, **237**, 109-114.
- ZHAO, N., HASHIDA, H., TAKAHASHI, N., MISUMI, Y. and SAKAKI, Y., 1995, High-density cDNA filter analysis: a novel approach for large scale quantitative analysis of gene expression. *Gene*, **156**, 207-213.
- ZHAO, X. J., NEWSOME, J. T. and CIHLAR, R. L., 1998, Up-regulation of two *Candida albicans* genes in the rat model of oral candidiasis detected by differential display. *Microbial Pathogenesis*, **25**, 121-129.
- ZIMMERMANN, C. R., ORR, W. C., LECLERC, R. F., BARNARD, C. and TIMBERLAKE, W. E., 1980, Molecular cloning and selection of genes regulated in *Aspergillus* development. *Cell*, **21**, 709-715.

Microarrays and Toxicology: The Advent of Toxicogenomics

Emile F. Nuwaysir,¹ Michael Bittner,² Jeffrey Trent,² J. Carl Barrett,¹ and Cynthia A. Afshari¹

¹Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina

²Laboratory of Cancer Genetics, National Human Genome Research Institute, Bethesda, Maryland

The availability of genome-scale DNA sequence information and reagents has radically altered life-science research. This revolution has led to the development of a new scientific subdiscipline derived from a combination of the fields of toxicology and genomics. This subdiscipline, termed toxicogenomics, is concerned with the identification of potential human and environmental toxicants, and their putative mechanisms of action, through the use of genomics resources. One such resource is DNA microarrays or "chips," which allow the monitoring of the expression levels of thousands of genes simultaneously. Here we propose a general method by which gene expression, as measured by cDNA microarrays, can be used as a highly sensitive and informative marker for toxicity. Our purpose is to acquaint the reader with the development and current state of microarray technology and to present our view of the usefulness of microarrays to the field of toxicology. *Mol. Carcinog.* 24:153-159, 1999. © 1999 Wiley-Liss, Inc.

Key words: toxicology; gene expression; animal bioassay

INTRODUCTION

Technological advancements combined with intensive DNA sequencing efforts have generated an enormous database of sequence information over the past decade. To date, more than 3 million sequences, totaling over 2.2 billion bases [1], are contained within the GenBank database, which includes the complete sequences of 19 different organisms [2]. The first complete sequence of a free-living organism, *Haemophilus influenzae*, was reported in 1995 [3] and was followed shortly thereafter by the first complete sequence of a eukaryote, *Saccharomyces cerevisiae* [4]. The development of dramatically improved sequencing methodologies promises that complete elucidation of the *Homo sapiens* DNA sequence is not far behind [5].

To exploit more fully the wealth of new sequence information, it was necessary to develop novel methods for the high-throughput or parallel monitoring of gene expression. Established methods such as northern blotting, RNase protection assays, S1 nuclease analysis, plaque hybridization, and slot blots do not provide sufficient throughput to effectively utilize the new genomics resources. Newer methods such as differential display [6], high-density filter hybridization [7,8], serial analysis of gene expression [9], and cDNA- and oligonucleotide-based microarray "chip" hybridization [10-12] are possible solutions to this bottleneck. It is our belief that the microarray approach, which allows the monitoring of expression levels of thousands of genes simultaneously, is a tool of unprecedented power for use in toxicology studies.

Almost without exception, gene expression is altered during toxicity, as either a direct or indirect result of toxicant exposure. The challenge facing toxicologists is to define, under a given set of experimental conditions, the characteristic and specific pattern of gene expression elicited by a given toxicant. Microarray technology offers an ideal platform for this type of analysis and could be the foundation for a fundamentally new approach to toxicology testing.

MICROARRAY DEVELOPMENT AND APPLICATIONS

cDNA Microarrays

In the past several years, numerous systems were developed for the construction of large-scale DNA arrays. All of these platforms are based on cDNAs or oligonucleotides immobilized to a solid support. In the cDNA approach, cDNA (or genomic) clones of interest are arrayed in a multi-well format and amplified by polymerase chain reaction. The products of this amplification, which are usually 500- to 2000-bp clones from the 3' regions of the genes of interest, are then spotted onto solid support by using high-speed robotics. By using this method, microarrays of up to 10 000 clones can be generated by spotting onto a glass substrate

*Correspondence to: Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, 111 Alexander Drive, Research Triangle Park, NC 27709.

Received 8 December 1998; Accepted 5 January 1999

Abbreviations: PAH, polycyclic aromatic hydrocarbon; NIEHS, National Institute of Environmental Health Sciences.

[13,14]. Sample detection for microarrays on glass involves the use of probes labeled with fluorescent or radioactive nucleotides.

Fluorescent cDNA probes are generated from control and test RNA samples in single-round reverse-transcription reactions in the presence of fluorescently tagged dUTP (e.g., Cy3-dUTP and Cy5-dUTP), which produces control and test products labeled with different fluorophores. The cDNAs generated from these two populations, collectively termed the "probe," are then mixed and hybridized to the array under a glass coverslip [10,11,15]. The fluorescent signal is detected by using a custom-designed scanning confocal microscope equipped with a motorized stage and lasers for fluor excitation [10,11,15]. The data are analyzed with custom digital image analysis software that determines for each DNA feature the ratio of fluor 1 to fluor 2, corrected for local background [16,17]. The strength of this approach lies in the ability to label RNAs from control and treated samples with different fluorescent nucleotides, allowing for the simultaneous hybridization and detection of both populations on one microarray. This method eliminates the need to control for hybridization between arrays. The research groups of Drs. Patrick Brown and Ron Davis at Stanford University spearheaded the effort to develop this approach, which has been successfully applied to studies of *Arabidopsis thaliana* RNA [10], yeast genomic DNA [15], tumorigenic versus non-tumorigenic human tumor cell lines [11], human T-cells [18], yeast RNA [19], and human inflammatory disease-related genes [20]. The most dramatic result of this effort was the first published account of gene expression of an entire genome, that of the yeast *Saccharomyces cerevisiae* [21].

In an alternative approach, large numbers of cDNA clones can be spotted onto a membrane support, albeit at a lower density [7,22]. This method is useful for expression profiling and large-scale screening and mapping of genomic or cDNA clones [7,22–24]. In expression profiling on filter membranes, two different membranes are used simultaneously for control and test RNA hybridizations, or a single membrane is stripped and reprobed. The signal is detected by using radioactive nucleotides and visualized by phosphorimager analysis or autoradiography. Numerous companies now sell such cDNA membranes and software to analyze the image data [25–27].

Oligonucleotide Microarrays

Oligonucleotide microarrays are constructed either by spotting prefabricated oligos on a glass support [13] or by the more elegant method of direct in situ oligo synthesis on the glass surface by photolithography [28–30]. The strength of this approach lies in its ability to discriminate DNA molecules based on single base-pair difference. This allows the application of this method to the fields of medical diagnos-

tics, pharmacogenetics, and sequencing by hybridization as well as gene-expression analysis.

Fabrication of oligonucleotide chips by photolithography is theoretically simple but technically complex [29,30]. The light from a high-intensity mercury lamp is directed through a photolithographic mask onto the silica surface, resulting in deprotection of the terminal nucleotides in the illuminated regions. The entire chip is then reacted with the desired free nucleotide, resulting in selected chain elongation. This process requires only $4n$ cycles (where n = oligonucleotide length in bases) to synthesize a vast number of unique oligos, the total number of which is limited only by the complexity of the photolithographic mask and the chip size [29,31,32].

Sample preparation involves the generation of double-stranded cDNA from cellular poly(A)⁺ RNA followed by antisense RNA synthesis in an in vitro transcription reaction with biotinylated or fluor-tagged nucleotides. The RNA probe is then fragmented to facilitate hybridization. If the indirect visualization method is used, the chips are incubated with fluor-linked streptavidin (e.g., phycoerythrin) after hybridization [12,33]. The signal is detected with a custom confocal scanner [34]. This method has been applied successfully to the mapping of genomic library clones [35], to de novo sequencing by hybridization [28,36], and to evolutionary sequence comparison of the *BRCA1* gene [37]. In addition, mutations in the cystic fibrosis [38] and *BRCA1* [39] gene products and polymorphisms in the human immunodeficiency virus-1 clade B protease gene [40] have been detected by this method. Oligonucleotide chips are also useful for expression monitoring [33] as has been demonstrated by the simultaneous evaluation of gene-expression patterns in nearly all open reading frames of the yeast strain *S. cerevisiae* [12]. More recently, oligonucleotide chips have been used to help identify single nucleotide polymorphisms in the human [41] and yeast [42] genomes.

THE USE OF MICROARRAYS IN TOXICOLOGY

Screening for Mechanism of Action

The field of toxicology uses numerous in vivo model systems, including the rat, mouse, and rabbit, to assess potential toxicity and these bioassays are the mainstay of toxicology testing. However, in the past several decades, a plethora of in vitro techniques have been developed to measure toxicity, many of which measure toxicant-induced DNA damage. Examples of these assays include the Ames test, the Syrian hamster embryo cell transformation assay, micronucleus assays, measurements of sister chromatid exchange and unscheduled DNA synthesis, and many others. Fundamental to all of these methods is the fact that toxicity is often preceded by, and results in, alterations in gene expression. In many cases, these changes in gene expression are a

far more sensitive, characteristic, and measurable endpoint than the toxicity itself. We therefore propose that a method based on measurements of the genome-wide gene expression pattern of an organism after toxicant exposure is fundamentally informative and complements the established methods described above.

We are developing a method by which toxicants can be identified and their putative mechanisms of action determined by using toxicant-induced gene expression profiles. In this method, in one or more defined model systems, dose and time-course parameters are established for a series of toxicants within a given prototypic class (e.g., polycyclic aromatic hydrocarbons (PAHs)). Cells are then treated with these agents at a fixed toxicity level (as measured by cell survival), RNA is harvested, and toxicant-induced gene expression changes are assessed by hybridization to a cDNA microarray chip (Figure 1). We have developed a custom DNA chip, called ToxChip v1.0, specifically for this purpose and will discuss it in more detail below. The changes in gene expression induced by the test agents in the model systems are analyzed, and the common set of changes unique to that class of toxicants, termed a toxicant signature, is determined.

This signature is derived by ranking across all experiments the gene-expression data based on rela-

tive fold induction or suppression of genes in treated samples versus untreated controls and selecting the most consistently different signals across the sample set. A different signature may be established for each prototypic toxicant class. Once the signatures are determined, gene-expression profiles induced by unknown agents in these same model systems can then be compared with the established signatures. A match assigns a putative mechanism of action to the test compound. Figure 2 illustrates this signature method for different types of oxidant stressors, PAHs, and peroxisome proliferators. In this example, the unknown compound in question had a gene-expression profile similar to that of the oxidant stressors in the database. We anticipate that this general method will also reveal cross talk between different pathways induced by a single agent (e.g., reveal that a compound has both PAH-like and oxidant-like properties). In the future, it may be necessary to distinguish very subtle differences between compounds within a very large sample set (e.g., thousands of highly similar structural isomers in a combinatorial chemistry library or peptide library). To generate these highly refined signatures, standard statistical clustering techniques or principal-component analysis can be used.

For the studies outlined in Figure 2, we developed the custom cDNA microarray chip ToxChip v1.0.

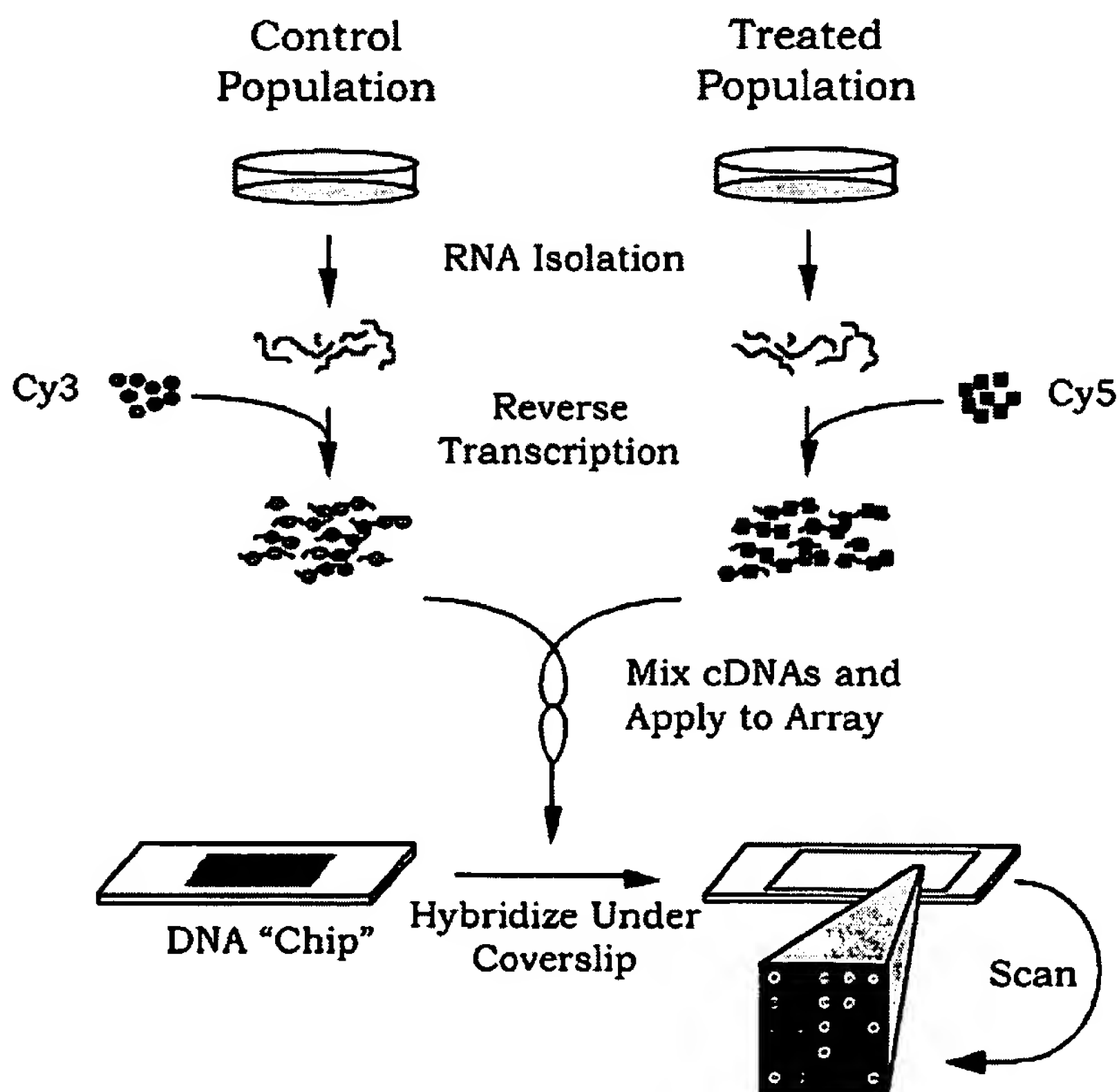


Figure 1. Simplified overview of the method for sample preparation and hybridization to cDNA microarrays. For illus-

trative purposes, samples derived from cell culture are depicted, although other sample types are amenable to this analysis.

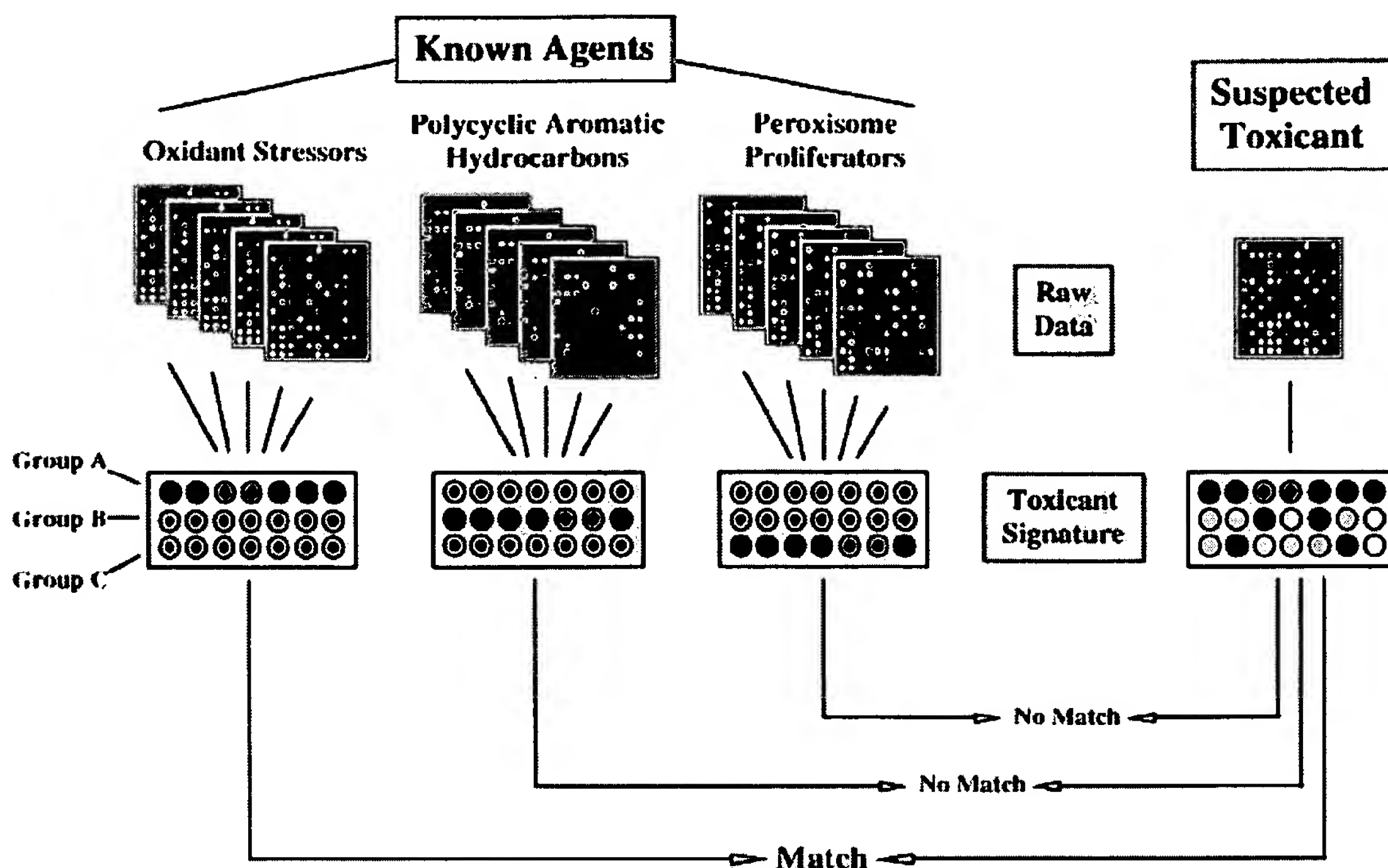


Figure 2. Schematic representation of the method for identification of a toxicant's mechanism of action. In this method, gene-expression data derived from exposure of model systems to known toxicants are analyzed, and a set of changes characteristic to that type of toxicant (termed the toxicant signature) is identified. As depicted, oxidant stressors produce

consistent changes in group A genes (indicated by red and green circles), but not group B or C genes (indicated by gray circles). The set of gene-expression changes elicited by the suspected toxicant is then compared with these characteristic patterns, and a putative mechanism of action is assigned to the unknown agent.

The 2090 human genes that comprise this subarray were selected for their well-documented involvement in basic cellular processes as well as their responses to different types of toxic insult. Included on this list are DNA replication and repair genes, apoptosis genes, and genes responsive to PAHs and dioxin-like compounds, peroxisome proliferators, estrogenic compounds, and oxidant stress. Some of the other categories of genes include transcription factors, oncogenes, tumor suppressor genes, cyclins, kinases, phosphatases, cell adhesion and motility genes, and homeobox genes. Also included in this group are 84 housekeeping genes, whose hybridization intensity is averaged and used for signal normalization of the other genes on the chip. To date, very few toxicants have been shown to have appreciable effects on the expression of these housekeeping genes. However, this housekeeping list will be revised if new data warrant the addition or deletion of a particular gene. Table 1 contains a general description of some of the different classes of genes that comprise ToxChip v1.0.

When a toxicant signature is determined, the genes within this signature are flagged within the database. When uncharacterized toxicants are then screened, the data can be quickly reformatted so that blocks of genes representing the different signatures

are displayed [11]. This facilitates rapid, visual interpretation of data. We are also developing ToxChip v2.0 and chips for other model systems, including rat, mouse, *Xenopus*, and yeast, for use in toxicology studies.

Animal Models in Toxicology Testing

The toxicology community relies heavily on the use of animals as model systems for toxicology testing. Unfortunately, these assays are inherently expensive, require large numbers of animals and take a long time to complete and analyze. Therefore, the National Institute of Environmental Health Sciences (NIEHS), the National Toxicology Program, and the toxicology community at large are committed to reducing the number of animals used, by developing more efficient and alternative testing methodologies. Although substantial progress has been made in the development of alternative methods, bioassays are still used for testing endpoints such as neurotoxicity, immunotoxicity, reproductive and developmental toxicology, and genetic toxicology. The rodent cancer bioassay is a particularly expensive and time-consuming assay, as it requires almost 4 yr, 1200 animals, and millions of dollars to execute and analyze [43]. In vitro experiments of the type outlined in Figure 2 might provide evidence that an unknown

Table 1. ToxChip v1.0: A Human cDNA Microarray Chip Designed to Detect Responses to Toxic Insult

Gene category	No. of genes on chip
Apoptosis	72
DNA replication and repair	99
Oxidative stress/redox homeostasis	90
Peroxisome proliferator responsive	22
Dioxin/PAH responsive	12
Estrogen responsive	63
Housekeeping	84
Oncogenes and tumor suppressor genes	76
Cell-cycle control	51
Transcription factors	131
Kinases	276
Phosphatases	88
Heat-shock proteins	23
Receptors	349
Cytochrome P450s	30

*This list is intended as a general guide. The gene categories are not unique, and some genes are listed in multiple categories.

agent is (or is not) responsible for eliciting a given biological response. This information would help to select a bioassay more specifically suited to the agent in question or perhaps suggest that a bioassay is not necessary, which would dramatically reduce cost, animal use, and time.

The addition of microarray techniques to standard bioassays may dramatically enhance the sensitivity and interpretability of the bioassay and possibly reduce its cost. Gene-expression signatures could be determined for various types of tissue-specific toxicants, and new compounds could be screened for these characteristic signatures, providing a rapid and sensitive *in vivo* test. Also, because gene expression is often exquisitely sensitive to low doses of a toxicant, the combination of gene-expression screening and the bioassay might allow the use of lower toxicant doses, which are more relevant to human exposure levels, and the use of fewer animals. In addition, gene-expression changes are normally measured in hours or days, not in the months to years required for tumor development. Furthermore, microarrays might be particularly useful for investigating the relationship between acute and chronic toxicity and identifying secondary effects of a given toxicant by studying the relationship between the duration of exposure to a toxicant and the gene-expression profile produced. Thus, a bioassay that incorporates gene-expression signatures with traditional endpoints might be substantially shorter, use more realistic dose regimens, and cost substantially less than the current assays do.

These considerations are also relevant for branches of toxicology not related to human health and not using rodents as model systems, such as aquatic toxicology and plant pathology. Bioassays based on the flathead minnow, *Daphnia*, and *Arabidopsis* could

also be improved by the addition of microarray analysis. The combination of microarrays with traditional bioassays might also be useful for investigating some of the more intractable problems in toxicology research, such as the effects of complex mixtures and the difficulties in cross-species extrapolation.

Exposure Assessment, Environmental Monitoring, and Drug Safety

The currently used methods for assessment of exposure to chemical toxicants are based on measurement of tissue toxin levels or on surrogate markers of toxicity, termed biomarkers (e.g., peripheral blood levels of hepatic enzymes or DNA adducts). Because gene expression is a sensitive endpoint, gene expression as measured with microarray technology may be useful as a new biomarker to more precisely identify hazards and to assess exposure. Similarly, microarrays could be used in an environmental-monitoring capacity to measure the effect of potential contaminants on the gene-expression profiles of resident organisms. In an analogous fashion, microarrays could be used to measure gene-expression endpoints in subjects in clinical trials. The combination of these gene-expression data and more established toxic endpoints in these trials could be used to define highly precise surrogates of safety.

Gene-expression profiles in samples from exposed individuals could be compared to the profiles of the same individuals before exposure. From this information, the nature of the toxic exposure can be determined or a relative clinical safety factor estimated. In the future it may also be possible to estimate not only the nature but the dose of the toxicant for a given exposure, based on relative gene-expression levels. This general approach may be particularly appropriate for occupational-health applications, in which unexposed and exposed samples from the same individuals may be obtainable. For example, a pilot study of gene expression in peripheral-blood lymphocytes of Polish coke-oven workers exposed to PAHs (and many other compounds) is under consideration at the NIEHS. An important consideration for these types of studies is that gene expression can be affected by numerous factors, including diet, health, and personal habits. To reduce the effects of these confounding factors, it may be necessary to compare pools of control samples with pools of treated samples. In the future it may be possible to compare exposed sample sets to a national database of human-expression data, thus eliminating the need to provide an unexposed sample from the same individual. Efforts to develop such a national gene-expression database are currently under way [44,45]. However, this national database approach will require a better understanding of genome-wide gene expression across the highly diverse human population and of the effects of environmental factors on this expression.

Alleles, Oligo Arrays, and Toxicogenetics

Gene sequences vary between individuals, and this variability can be a causative factor in human diseases of environmental origin [46,47]. A new area of toxicology, termed toxicogenetics, was recently developed to study the relationship between genetic variability and toxicant susceptibility. This field is not the subject of this discussion, but it is worthwhile to note that the ability of oligonucleotide arrays to discriminate DNA molecules based on single base-pair differences makes these arrays uniquely useful for this type of analysis. Recent reports demonstrated the feasibility of this approach [41,42]. The NIEHS has initiated the Environmental Genome Project to identify common sequence polymorphisms in 200 genes thought to be involved in environmental diseases [48]. In a pilot study on the feasibility of this application to the Environmental Genome Project, oligonucleotide arrays will be used to resequence 20 candidate genes. This toxicogenetic approach promises to dramatically improve our understanding of interindividual variability in disease susceptibility.

FUTURE PRIORITIES

There are many issues that must be addressed before the full potential of microarrays in toxicology research can be realized. Among these are model system selection, dose selection, and the temporal nature of gene expression. In other words, in which species, at what dose, and at what time do we look for toxicant-induced gene expression? If human samples are analyzed, how variable is global gene expression between individuals, before and after toxicant exposure? What are the effects of age, diet, and other factors on this expression? Experience, in the form of large data sets of toxicant exposures, will answer these questions.

One of the most pressing issues for array scientists is the construction of a national public database (linked to the existing public databases) to serve as a repository for gene-expression data. This relational database must be made available for public use, and researchers must be encouraged to submit their expression data so that others may view and query the information. Researchers at the National Institutes of Health have made laudable progress in developing the first generation of such a database [44,45]. In addition, improved statistical methods for gene clustering and pattern recognition are needed to analyze the data in such a public database.

The proliferation of different platforms and methods for microarray hybridizations will improve sample handling and data collection and analysis and reduce costs. However, the variety of microarray methods available will create problems of data compatibility between platforms. In addition, the near-infinite variety of experimental conditions under

which data will be collected by different laboratories will make large-scale data analysis extremely difficult. To help circumvent these future problems, a set of standards to be included on all platforms should be established. These standards would facilitate data entry into the national database and serve as reference points for cross-platform and inter-laboratory data analysis.

Many issues remain to be resolved, but it is clear that new molecular techniques such as microarray hybridization will have a dramatic impact on toxicology research. In the future, the information gathered from microarray-based hybridization experiments will form the basis for an improved method to assess the impact of chemicals on human and environmental health.

ACKNOWLEDGMENTS

The authors would like to thank Drs. Robert Maronpot, George Lucier, Scott Masten, Nigel Walker, Raymond Tennant, and Ms. Theodora Deverenux for critical review of this manuscript. EFN was supported in part by NIEHS Training Grant #ES07017-24.

REFERENCES

1. <http://www.ncbi.nlm.nih.gov/Web/Genbank/index.html>
2. <http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html>
3. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 1995;269:496-512.
4. Goffeau A, Barrell BG, Bussey H, et al. Life with 6000 genes. *Science* 1996;274:546, 563-567.
5. <http://www.perkin-elmer.com/press/prc5448.html>
6. Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* 1992;257:967-971.
7. Pietu G, Alibert O, Guichard V, et al. Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. *Genome Res* 1996;6:492-503.
8. Zhao ND, Hashida H, Takahashi N, Misumi Y, Sakaki Y. High-density cDNA filter analysis—A novel approach for large-scale, quantitative analysis of gene expression. *Gene* 1995;156:207-213.
9. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science* 1995;270:484-487.
10. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene-expression patterns with a complementary DNA microarray. *Science* 1995;270:467-470.
11. DeRisi J, Penland L, Brown PO, et al. use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* 1996;14:457-460.
12. Wodicka L, Dong HL, Mittmann M, Ho MH, Lockhart DJ. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat Biotechnol* 1997;15:1359-1367.
13. Marshall A, Hodgson J. DNA chips: An array of possibilities. *Nat Biotechnol* 1998;16:27-31.
14. <http://www.synteni.com>
15. Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res* 1996;6:639-645.
16. Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Biomedical Optics* 1997;2:364-374.
17. Khan J, Simon R, Bittner M, et al. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res* 1998;58:5009-5013.
18. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci USA* 1996; 93:10614-10619.

19. Lashkari DA, DeRisi JL, McCusker JH, et al. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc Natl Acad Sci USA* 1997;94:13057-13062.
20. Heller RA, Schena M, Chai A, et al. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc Natl Acad Sci USA* 1997;94:2150-2155.
21. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 1997;278:680-686.
22. Drmanac S, Stavropoulos NA, Labat I, et al. Gene-representing cDNA clusters defined by hybridization of 57,419 clones from infant brain libraries with short oligonucleotide probes. *Genomics* 1996;37:29-40.
23. Milosavljevic A, Savkovic S, Crkvenjakov R, et al. DNA sequence recognition by hybridization to short oligomers: Experimental verification of the method on the *E. coli* genome. *Genomics* 1996;37:77-86.
24. Drmanac S, Drmanac R. Processing of cDNA and genomic kilobase-size clones for massive screening, mapping and sequencing by hybridization. *Biotechniques* 1994;17:328-329, 332-336.
25. <http://www.resgen.com/>
26. <http://www.genomesystems.com/>
27. <http://www.clontech.com/>
28. Pease AC, Solas DA, Fodor SPA. Parallel synthesis of spatially addressable oligonucleotide probe matrices. *Abstracts of Papers of the American Chemical Society* 1992;203:34.
29. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SPA. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci USA* 1994;91:5022-5026.
30. Fodor SPA, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. *Science* 1991;251:767-773.
31. McGall G, Labadie J, Brock P, Wallraff G, Nguyen T, Hinsberg W. Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists. *Proc Natl Acad Sci USA* 1996;93:13555-13560.
32. Lipshutz RJ, Morris D, Chee M, et al. Using oligonucleotide probe arrays to access genetic diversity. *Biotechniques* 1995;19:442-447.
33. Lockhart DJ, Dong HL, Byrne MC, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 1996;14:1675-1680.
34. <http://www.mdyn.com/>
35. Sapolsky RJ, Lipshutz RJ. Mapping genomic library clones using oligonucleotide arrays. *Genomics* 1996;33:445-456.
36. Chee M, Yang R, Hubbell E, et al. Accessing genetic information with high-density DNA arrays. *Science* 1996;274:610-614.
37. Hacia JG, Makalowski W, Edgemon K, et al. Evolutionary sequence comparisons using high-density oligonucleotide arrays. *Nat Genet* 1998;18:155-158.
38. Cronin MT, Fucini RV, Kim SM, Masino RS, Wespi RM, Miyada CG. Cystic fibrosis mutation detection by hybridization to light-generated DNA probe arrays. *Hum Mutat* 1996;7:244-255.
39. Hacia JG, Brody LC, Chee MS, Fodor SPA, Collins FS. Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis. *Nat Genet* 1996;14:441-447.
40. Kozal MJ, Shah N, Shen NP, et al. Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. *Nat Med* 1996;2:753-759.
41. Wang DG, Fan JB, Siao CJ, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 1998;280:1077-1082.
42. Winzeler EA, Richards DR, Conway AR, et al. Direct allelic variation scanning of the yeast genome. *Science* 1998;281:1194-1197.
43. Chhabra RS, Huff JE, Schwetz BS, Selkirk J. An overview of prechronic and chronic toxicity carcinogenicity experimental-study designs and criteria used by the National Toxicology Program. *Environ Health Perspect* 1990;86:313-321.
44. Ermolaeva O, Rastogi M, Pruitt KD, et al. Data management and analysis for gene expression arrays. *Nat Genet* 1998;20:19-23.
45. <http://www.nhgri.nih.gov/DIR/LCG/15K/HTML/dbase.html>
46. Samson M, Libert F, Doranz BJ, et al. Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature* 1996;382:722-725.
47. Bell DA, Taylor JA, Paulson DF, Robertson CN, Mohler JL, Lucier GW. Genetic risk and carcinogen exposure—A common inherited defect of the carcinogen-metabolism gene glutathione-S-transferase M1 (*Gstm1*) that increases susceptibility to bladder cancer. *J Natl Cancer Inst* 1993;85:1159-1164.
48. <http://www.niehs.nih.gov/envgenom/home.html>

Pharmaceutical Proteomics

SANDRA STEINER^a AND N. LEIGH ANDERSON

Large Scale Proteomics Corporation, Rockville, Maryland, USA

ABSTRACT: Genomics and proteomics are today well established in drug discovery and, in combination with combinatorial chemistry and high-throughput screening, are helping to bring forward an unprecedented number of potential lead compounds. To avoid the generation of bottlenecks downstream in drug development, increasing pressure is arising to integrate these technologies into the development environment. Proteomics has demonstrated proof-of-concept in toxicology as shown by a number of successful applications in mechanistic toxicology and lead selection. The "technology wave" is now starting to impact the clinical phase of drug development. Expected benefits are optimized clinical trials based on the availability of biologically relevant markers of drug efficacy and safety.

INTRODUCTION

In recent years, a number of technology developments have had a profound impact on the drug discovery process. The exponential growth in genomics and proteomics capabilities and the subsequent generation of large amounts of novel information have facilitated an unprecedented number of potential new drug targets. Advances in combinatorial chemistry resulting in a nearly unlimited availability of compound libraries, in combination with constant improvements in high-throughput screening (HTS) technologies, have resulted in unprecedented numbers of potential lead compounds. A consequence of this technology-driven acceleration of drug discovery is the creation of bottlenecks downstream in drug development. Thus, increasing pressure is currently arising to integrate genomics and proteomics approaches in the development environment to introduce a similar boost to drug development. The implementation of proteomics in safety assessment has actually advanced beyond the proof-of-concept stage as demonstrated by a significant number of studies performed by our laboratory and by other investigators.

PROTEOME PROFILING TO OBTAIN INSIGHTS INTO MECHANISMS AND PATHWAYS OF TOXICITY

Insights into molecular mechanisms and knowledge of biological pathways involved in toxic responses are an asset for the interpretation of adverse drug effects and contribute to an accurate risk assessment for humans. There is abundant evidence that links can be established between the up- or downregulation of specific

^aAddress for correspondence: S. Steiner, Large Scale Biology Corporation, 9620 Medical Center Drive, Rockville, MD 20850-3338. Voice: 301-424-5989; fax: 301-762-4892. sandra.steiner@lsbc.com

pathways and the morphological manifestation of toxic endpoints, as shown for example with the chemically diverse group of peroxisome proliferating compounds.^{1,2} We have studied the liver effects of a set of strong peroxisome proliferators (PP; agonists of the nuclear receptor PPAR α) and a nonproliferator (an analogue of one of the PPs tested having similar pharmacological potency) in mice and demonstrated three distinct proteome signature patterns.³ The patterns that could be distinguished corresponded to (i) PPAR α agonist activity, (ii) animal age (30 days difference in age at sacrifice between 5- and 35-day treatment groups), and (iii) action of the non-PP compound by a different mechanism. These results provided support for a unified receptor-based mechanism controlling the main PP response, but demonstrate that individual responsive genes can show quite different dose-response curves. More than 100 proteins showed significant changes following PP treatment, and multiple sensitive markers were identified in this project. Based on the assumption that impairments of biological pathways (as visualized in proteome signature patterns) precede potential morphological manifestations, proteins showing coherent changes following treatment with PPs are likely to be early markers for stimuli that eventually result in proliferation of peroxisomes.

Proteomics was the key to new insights into the molecular mechanisms involved in cyclosporine A (CsA) nephrotoxicity.^{4,5} The use of CsA, a potent immunosuppressant, is limited by its kidney toxicity. In kidney proteome patterns from CsA-treated rats, we showed a profound downregulation of the calcium binding protein, calbindin D28, an intracellular calcium buffer and transport protein. Its near absence in the kidneys of CsA-treated animals provides an explanation for the accumulation of calcium in the tubules and consequent tubular toxicity. A subsequent SAR study showed that the downregulation of calbindin was closely associated with immunosuppressant activity⁶ and that its downregulation also occurred in humans showing CsA-related nephrotoxicity. Prior to the proteome study, the link in relationship between CsA kidney toxicity and calbindin D28 downregulation was not known.

PROTEOME PROFILING AND LEAD SELECTION

The selection of lead compounds is a critical step in the drug development process with profound downstream consequences. The selection pressure favors compounds with a wide therapeutic window, as defined by high pharmacological potency in combination with low toxicity. Estimation of therapeutic windows in light of the limited data sets typically available for compounds at the late discovery/early development phase is often similar to guesswork. Thus, information residing in proteome profiles produced from lead candidates can be essential to support compound prioritization decisions.

We used proteome profiling as a basis to select protein markers linked to compound efficacy or toxicity and demonstrated the possibility of using these markers for lead prioritization.⁷ A compound studied in this context was SDZ PGU 693, a hypoglycemic agent found to induce hepatocellular hypertrophy in the rat. Liver proteome profiling of treated rats showed the induction of several microsomal proteins, including NADPH cytochrome P-450 reductase and cytochrome b5, indicative of microsomal proliferation and induction of the P-450 enzyme system causing hepatocellular hypertrophy. Decreases were evident in a series of mitochondrial proteins

such as F₁ATPase-alpha subunit and cytosolic liver fatty acid binding protein suggesting a downregulation of the mitochondrial liver fatty acid metabolism, likely reflecting the pharmacological action of the compound. These data demonstrate that the liver proteome of treated rats revealed protein markers indicative for both SDZ PGU efficacy and toxicity. Markers for both endpoints were selected from these profiles and high-throughput protein assays set up to screen follow-up compounds to assess an estimate of their therapeutic window.

Similarly, efficacy and toxicity markers for "statin"-class cholesterol-lowering compounds were selected using proteome analysis. We found that agents acting to alter blood cholesterol (e.g., the statin HMG-CoA reductase inhibitors, cholestyramine, and high-cholesterol diets) change the abundances of several proteins in rat liver.⁸ Most strongly affected is a protein identified as HMG-CoA synthase, a critical enzyme of the cholesterol synthesis pathway. Based on these data, HMG-CoA synthase can be used as an intracellular reporter of the pathway's performance. The statins, but not cholestyramine, also induce a peroxisomal enoyl hydratase-like protein, which was identified as a sensitive marker for peroxisome proliferation (representing in this case an undesirable side effect). The ratio between these efficacy and toxicity markers is different for different members of the marketed statin drugs, indicating their likely pharmacological inequivalence.

A study that we performed with etomoxir resulted in the observation that drug-induced liver steatosis is linked to the expression of a protein considered to be adipocyte-specific.⁹ The liver toxicity of etomoxir, an inhibitor of carnitine palmitoyl transferase developed as a potential antidiabetic, was found to involve induction of the "adipocyte differentiation-related protein" (ADRP). This protein appears to be associated with lipid droplets that accumulate in the hepatocytes following treatment with etomoxir and demonstrates that at least one gene product previously considered specific to the adipocyte is induced in hepatocytes by the drug. ADRP may serve as an early toxicity marker for impaired lipid metabolism and lipid accumulation in liver.

CONCLUSIONS AND PERSPECTIVES

The basic methodology of safety evaluation has changed little during the past decades. Toxicity in laboratory animals has been evaluated primarily by using hematological, clinical chemistry, and histological parameters as indicators of organ damage. There is an emerging number of studies proving the unique value of global approaches such as proteome analysis to obtain crucial insights into mechanisms of toxicity and to pave the way towards predictive toxicology. All these data justify the expectation that the integration of proteomics and genomics into state-of-the-art toxicology will significantly advance the safety assessment of new drugs.

It becomes obvious that the impact of these new technologies will continue downstream into the development process and will extend from the preclinical to the clinic phase. As a consequence, former bottlenecks may disappear and new ones may appear and will need to be addressed. One of these is apparent today and arises from the necessity to obtain proof-of-concept of new drug candidates in humans as early as possible. Phase II and III clinical trials are the most expensive studies in the drug development process and failures at this stage may be fatal (both to the drug and,

unfortunately, occasionally to patients). The possibility to perform early proof-of-concept trials is greatly dependent on the availability of appropriate markers for drug efficacy and safety. Markers of this kind need to be easily accessible in biological fluids such as serum or urine to guarantee the possibility for frequent monitoring (which makes them likely to be proteins or peptides). The detection and validation of such markers are far from trivial tasks, and technologies such as proteomics, allowing one to mine serum, urine, and other biological fluids for relevant markers, are expected to be key players in this effort. Pressure is also growing to improve the selection of patient populations for clinical trials. Genomics and proteomics are expected to occupy central roles in the efforts to stratify patient populations. Finally, a vast amount of biologically highly relevant data will cycle back from the development to the discovery end, generating a continuous impulse that will stimulate the drug pipeline.

REFERENCES

1. REDDY, J.K. & D.L. AZARNOFF. 1980. Hypolipidemic hepatic peroxisome proliferators form a novel class of chemical carcinogens. *Nature* 283: 397-398.
2. DREYER, C., G. KREY, H. KELLER, F. GIVEL, G. HELFTENBEIN & W. WAHL. 1992. Control of the peroxisomal β -oxidation pathway by a novel family of nuclear hormone receptors. *Cell* 68: 879-887.
3. ANDERSON, N.L., R. ESQUER-BLASCO, F. RICHARDSON, P. FOXWORTHY & P. EACHO. 1996. The effects of peroxisome proliferators on protein abundances in mouse liver. *Toxicol. Appl. Pharmacol.* 137: 75-89.
4. AICHER, L., D. WAHL, A. ARCE, O. GRENET & S. STEINER. 1998. New insights into cyclosporine A nephrotoxicity by proteome analysis. *Electrophoresis* 19: 1998-2003.
5. STEINER, S., L. AICHER, J. RAYMACKERS, L. MEHEUS, R. ESQUER-BLASCO, N.L. ANDERSON & A. CORDIER. 1996. Cyclosporine A mediated decrease in the rat renal calcium binding protein calbindin-D 28kDa. *Biochem. Pharmacol.* 51: 253-258.
6. AICHER, L., G. MEIER, A. NORCROSS, J. JAKUBOWSKI, M.C. VARELA, A. CORDIER & S. STEINER. 1997. Decrease in kidney calbindin-D as a possible mechanism mediating CsA and FK-506-induced calciuria and tubular mineralization. *Biochem. Pharmacol.* 53: 723-731.
7. ARCE, A., L. AICHER, D. WAHL, R. ESQUER-BLASCO, N.L. ANDERSON, A. CORDIER & S. STEINER. 1998. Changes in the liver proteome of female Wistar rats treated with the hypoglycemic agent SDZ PGU 693. *Life Sci.* 63: 2243-2250.
8. ANDERSON, N.L., R. ESQUER-BLASCO, J.-P. HOFMANN & N.G. ANDERSON. 1991. A two-dimensional gel database of rat liver proteins useful in gene regulation and drug effects studies. *Electrophoresis* 12: 907-930.
9. STEINER, S., D. WAHL, B.L.K. MANGOLD, R. ROBISON, J. RAYMACKERS, L. MEHEUS, N.L. ANDERSON & A. CORDIER. 1996. Induction of the adipose differentiation-related protein in liver of etomoxir treated rats. *BBRC* 218: 777-782.

such as F₁ATPase-alpha subunit and cytosolic liver fatty acid binding protein suggesting a downregulation of the mitochondrial liver fatty acid metabolism, likely reflecting the pharmacological action of the compound. These data demonstrate that the liver proteome of treated rats revealed protein markers indicative for both SDZ PGU efficacy and toxicity. Markers for both endpoints were selected from these profiles and high-throughput protein assays set up to screen follow-up compounds to assess an estimate of their therapeutic window.

Similarly, efficacy and toxicity markers for "statin"-class cholesterol-lowering compounds were selected using proteome analysis. We found that agents acting to alter blood cholesterol (e.g., the statin HMG-CoA reductase inhibitors, cholestyramine, and high-cholesterol diets) change the abundances of several proteins in rat liver.⁸ Most strongly affected is a protein identified as HMG-CoA synthase, a critical enzyme of the cholesterol synthesis pathway. Based on these data, HMG-CoA synthase can be used as an intracellular reporter of the pathway's performance. The statins, but not cholestyramine, also induce a peroxisomal enoyl hydratase-like protein, which was identified as a sensitive marker for peroxisome proliferation (representing in this case an undesirable side effect). The ratio between these efficacy and toxicity markers is different for different members of the marketed statin drugs, indicating their likely pharmacological inequivalence.

A study that we performed with etomoxir resulted in the observation that drug-induced liver steatosis is linked to the expression of a protein considered to be adipocyte-specific.⁹ The liver toxicity of etomoxir, an inhibitor of carnitine palmitoyl transferase developed as a potential antidiabetic, was found to involve induction of the "adipocyte differentiation-related protein" (ADRP). This protein appears to be associated with lipid droplets that accumulate in the hepatocytes following treatment with etomoxir and demonstrates that at least one gene product previously considered specific to the adipocyte is induced in hepatocytes by the drug. ADRP may serve as an early toxicity marker for impaired lipid metabolism and lipid accumulation in liver.

CONCLUSIONS AND PERSPECTIVES

The basic methodology of safety evaluation has changed little during the past decades. Toxicity in laboratory animals has been evaluated primarily by using hematological, clinical chemistry, and histological parameters as indicators of organ damage. There is an emerging number of studies proving the unique value of global approaches such as proteome analysis to obtain crucial insights into mechanisms of toxicity and to pave the way towards predictive toxicology. All these data justify the expectation that the integration of proteomics and genomics into state-of-the-art toxicology will significantly advance the safety assessment of new drugs.

It becomes obvious that the impact of these new technologies will continue downstream into the development process and will extend from the preclinical to the clinic phase. As a consequence, former bottlenecks may disappear and new ones may appear and will need to be addressed. One of these is apparent today and arises from the necessity to obtain proof-of-concept of new drug candidates in humans as early as possible. Phase II and III clinical trials are the most expensive studies in the drug development process and failures at this stage may be fatal (both to the drug and,

unfortunately, occasionally to patients). The possibility to perform early proof-of-concept trials is greatly dependent on the availability of appropriate markers for drug efficacy and safety. Markers of this kind need to be easily accessible in biological fluids such as serum or urine to guarantee the possibility for frequent monitoring (which makes them likely to be proteins or peptides). The detection and validation of such markers are far from trivial tasks, and technologies such as proteomics, allowing one to mine serum, urine, and other biological fluids for relevant markers, are expected to be key players in this effort. Pressure is also growing to improve the selection of patient populations for clinical trials. Genomics and proteomics are expected to occupy central roles in the efforts to stratify patient populations. Finally, a vast amount of biologically highly relevant data will cycle back from the development to the discovery end, generating a continuous impulse that will stimulate the drug pipeline.

REFERENCES

1. REDDY, J.K. & D.L. AZARNOFF. 1980. Hypolipidemic hepatic peroxisome proliferators form a novel class of chemical carcinogens. *Nature* 283: 397-398.
2. DREYER, C., G. KREY, H. KELLER, F. GIVEL, G. HELFTENBEIN & W. WAHL. 1992. Control of the peroxisomal β -oxidation pathway by a novel family of nuclear hormone receptors. *Cell* 68: 879-887.
3. ANDERSON, N.L., R. ESQUER-BLASCO, F. RICHARDSON, P. FOXWORTHY & P. BACHO. 1996. The effects of peroxisome proliferators on protein abundances in mouse liver. *Toxicol. Appl. Pharmacol.* 137: 75-89.
4. AICHER, L., D. WAHL, A. ARCE, O. GRENET & S. STEINER. 1998. New insights into cyclosporine A nephrotoxicity by proteome analysis. *Electrophoresis* 19: 1998-2003.
5. STEINER, S., L. AICHER, J. RAYMACKERS, L. MEHEUS, R. ESQUER-BLASCO, N.L. ANDERSON & A. CORDIER. 1996. Cyclosporine A mediated decrease in the rat renal calcium binding protein calbindin-D 28kDa. *Biochem. Pharmacol.* 51: 253-258.
6. AICHER, L., G. MEIER, A. NORCROSS, J. JAKUBOWSKI, M.C. VARELA, A. CORDIER & S. STEINER. 1997. Decrease in kidney calbindin-D as a possible mechanism mediating CsA and FK-506-induced calciuria and tubular mineralization. *Biochem. Pharmacol.* 53: 723-731.
7. ARCE, A., L. AICHER, D. WAHL, R. ESQUER-BLASCO, N.L. ANDERSON, A. CORDIER & S. STEINER. 1998. Changes in the liver proteome of female Wistar rats treated with the hypoglycemic agent SDZ PGU 693. *Life Sci.* 63: 2243-2250.
8. ANDERSON, N.L., R. ESQUER-BLASCO, J.-P. HOFMANN & N.G. ANDERSON. 1991. A two-dimensional gel database of rat liver proteins useful in gene regulation and drug effects studies. *Electrophoresis* 12: 907-930.
9. STEINER, S., D. WAHL, B.L.K. MANGOLD, R. ROBISON, J. RAYMACKERS, L. MEHEUS, N.L. ANDERSON & A. CORDIER. 1996. Induction of the adipose differentiation-related protein in liver of etomoxir treated rats. *BBRC* 218: 777-782.

Subject: RE: [Fwd: Toxicology Chip]**Date: Mon. 3 Jul 2000 08:09:45 -0400****From: "Afshari.Cynthia" <afshari@niehs.nih.gov>****To: "Diana Hamlet-Cox" <dianahc@incyte.com>**

You can see the list of clones that we have on our 12K chip at
<http://manuel.niehs.nih.gov/maps/guest/clonesrch.cfm>

We selected a subset of genes (2000K) that we believed critical to tox response and basic cellular processes and added a set of clones and ESTs to this. We have included a set of control genes (80+) that were selected by the NHGRI because they did not change across a large set of array experiments. However, we have found that some of these genes change significantly after tox treatments and are in the process of looking at the variation of each of these 80+ genes across our experiments.

Our chips are constantly changing and being updated and we hope that our data will lead us to what the toxchip should really be.

I hope this answers your question.

Cindy Afshari

> -----

> From: Diana Hamlet-Cox
 > Sent: Monday, June 26, 2000 8:52 PM
 > To: afshari@niehs.nih.gov
 > Subject: [Fwd: Toxicology Chip]

> Dear Dr. Afshari,

> Since I have not yet had a response from Bill Grigg, perhaps he was not
 > the right person to contact.

> Can you help me in this matter? I don't need to know the sequences,
 > necessarily, but I would like very much to know what types of sequences
 > are being used, e.g., GPCRs (more specific?), ion channels, etc.

> Diana Hamlet-Cox

> ----- Original Message -----

> Subject: Toxicology Chip
 > Date: Mon, 19 Jun 2000 18:31:48 -0700
 > From: Diana Hamlet-Cox <dianahc@incyte.com>
 > Organization: Incyte Pharmaceuticals
 > To: grigg@niehs.nih.gov

> Dear Colleague:

> I am doing literature research on the use of expressed genes as
 > pharmacotoxicology markers, and found the Press Release dated February
 > 29, 2000 regarding the work of the NIEHS in this area. I would like to
 > know if there is a resource I can access (or you could provide?) that
 > would give me a list of the 12,000 genes that are on your Human ToxChip
 > Microarray. In particular, I am interested in the criteria used to
 > select sequences for the ToxChip, including any control sequences
 > included in the microarray.

> Thank you for your assistance in this request.

> Diana Hamlet-Cox, Ph.D.
 > Incyte Genomics, Inc.

> --

> =====

> This email message is for the sole use of the intended recipient s and
> may contain confidential and privileged information subject to
> attorney-client privilege. Any unauthorized review, use, disclosure or
> distribution is prohibited. If you are not the intended recipient,
> please contact the sender by reply email and destroy all copies of the
> original message.
> =====
>
>

Proteomics: a major new technology for the drug discovery process

Martin J. Page, Bob Amess, Christian Rohlff, Colin Stubberfield and Raj Parekh

Proteomics is a new enabling technology that is being integrated into the drug discovery process. This will facilitate the systematic analysis of proteins across any biological system or disease, forwarding new targets and information on mode of action, toxicology and surrogate markers. Proteomics is highly complementary to genomic approaches in the drug discovery process and, for the first time, offers scientists the ability to integrate information from the genome, expressed mRNAs, their respective proteins and subcellular localization. It is expected that this will lead to important new insights into disease mechanisms and improved drug discovery strategies to produce novel therapeutics.

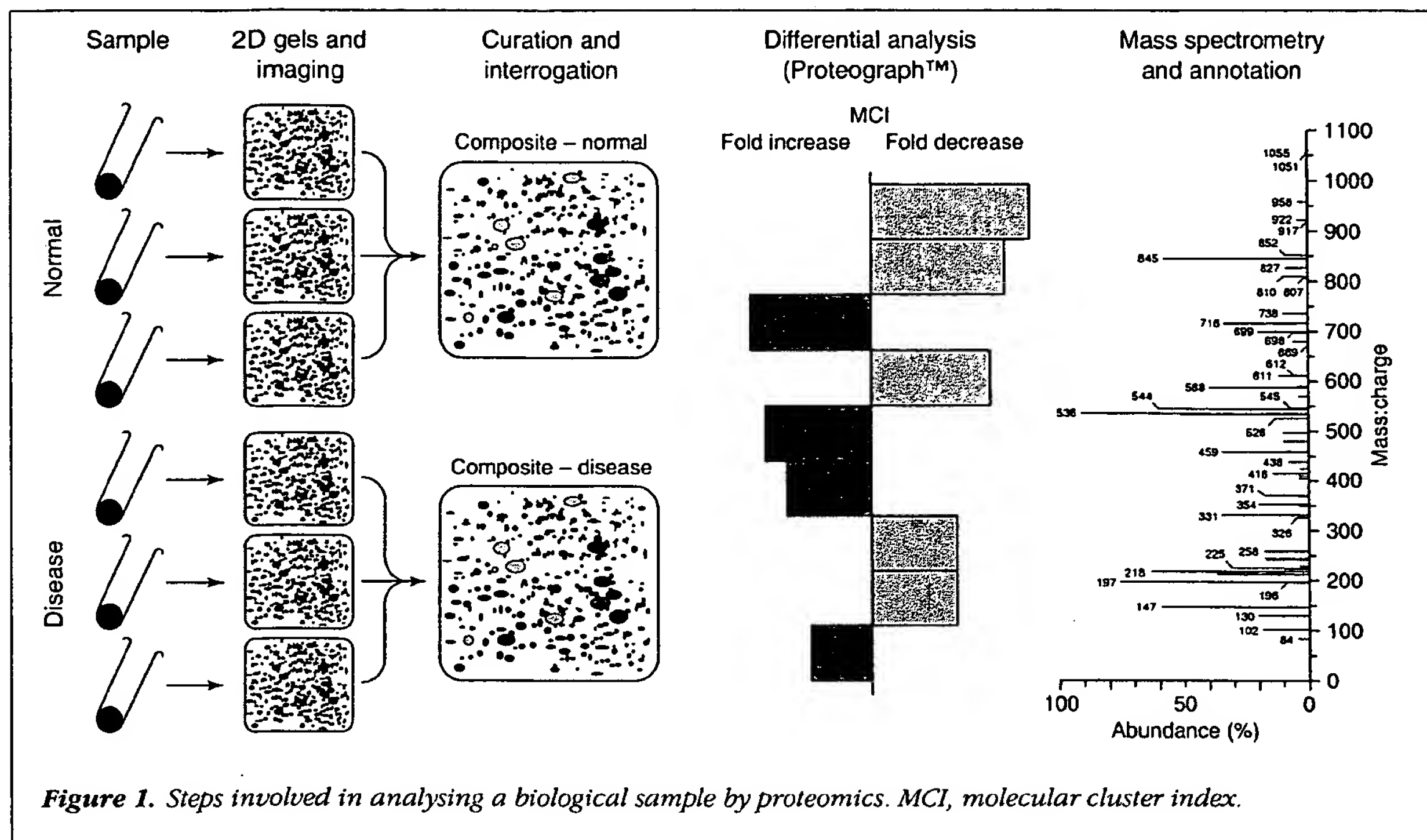
Among the major pharmaceutical and biotechnology companies, it is clearly recognized that the business of modern drug discovery is a highly competitive process. All of the many steps involved are inherently complex, and each can involve a high risk of attrition. The players in this business strive continuously to optimize and streamline the process; each seeking to gain an advantage at every step by attempting to make informed decisions at the earliest stage possible. The desired outcome is to accelerate as many key activities in the drug discovery process as possible. This should pro-

duce a new generation of robust drugs that offer a high probability of success and reach the clinic and market ahead of the competition.

There has been noticeable emphasis over recent years for companies to aggressively review and refine their strategies to discover new drugs. Central to this has been the introduction and implementation of cutting-edge technologies. Most, if not all, companies have now integrated key technology platforms that incorporate genomics, mRNA expression analysis, relational databases, high-throughput robotics, combinatorial chemistry and powerful bioinformatics. Although it is still early days to quantify the real impact of these platforms in clinical and commercial terms, expectations are high, and it is widely accepted that significant benefits will be forthcoming. This is largely based on data obtained during preclinical studies where the genomic^{1,2} and microarray^{3,4} technologies have already proved their value.

However, there are several noteworthy outcomes that result from this. Many comments are voiced that scientists armed with these technologies are now commonly faced with data overload. Thus, in some instances, rather than facilitating the decision process, the accumulation of more complex data points, many with unknown consequences, can seem to hinder the process. Also, most drug companies have simultaneously incorporated very similar components of the new technology platforms, the consequence being that it is becoming difficult yet again to determine where a clear competitive advantage will arise. Finally, in recent years, largely as a result of the accessibility of the technologies, there has been an overwhelming emphasis placed on genomic and mRNA data rather than on protein

Martin J. Page*, Bob Amess, Christian Rohlff, Colin Stubberfield and Raj Parekh, Oxford GlycoSciences, 10 The Quadrant, Abingdon Science Park, Abingdon, Oxfordshire, UK OX14 3YS. *tel: +44 1235 543277, fax: +44 1235 543283, e-mail: martin.page@ogs.co.uk



analysis. It is important to remember that proteins dictate biological phenotype – whether it is normal or diseased – and are the direct targets for most drugs.

Proteomics: new technology for the analysis of proteins

It is now timely to recognize that complementary technology in the form of high-throughput analysis of the total protein repertoire of chosen biological samples, namely proteomics, is poised to add a new and important dimension to drug discovery. In a similar fashion to genomics, which aims to profile every gene expressed in a cell, proteomics seeks to profile every protein that is expressed⁵⁻⁷. However, there is added information, since proteomics can also be used to identify the post-translational modifications of proteins⁸, which can have profound effects on biological function, and their cellular localization. Importantly, proteomics is a technology that integrates the significant advances in two-dimensional (2D) electrophoretic separation of proteins, mass spectrometry and bioinformatics. With these advances it is now possible to consistently derive proteomes that are highly reproducible and suitable for interrogation using advanced bioinformatic tools.

There are many variations whereby different laboratories operate proteomics. For the purpose of this review, the

process used at Oxford GlycoSciences (OGS), which uses an industrial-scale operation that is integral to its drug discovery work, will be described. The individual steps of this process, where up to 1000 2D gels can be run and analysed per week, are summarized in Fig. 1. The incoming samples are bar coded and all information relevant to the sample is logged into a Laboratory Information Management System (LIMS) database. There can be a wide range in the type of samples processed, as applicable to individual steps in the drug discovery pipeline, and these will be mentioned later. The samples are separated according to their charge (pI) in the first dimension, using isoelectric focusing, followed by size (MW) using SDS-PAGE in the second dimension. Many modifications have been made to these steps to improve handling, throughput and reproducibility. The separated proteins are then stained with fluorescent dyes which are significantly more sensitive in detection than standard silver methods and have a broader dynamic range. The image of the displayed proteins obtained is referred to as the proteome, and is digitally scanned into databases using proprietary software called ROSETTA™. The images are subsequently curated, which begins with the removal of any artefacts, cropping and the placement of pI/MW landmarks. The images from replicate images are then aligned and matched to one

another to generate a synthetic composite image. This is an important step, as the proteome is a dynamic situation, and it captures the biological variation that occurs, such that even orphan proteins are still incorporated into the analysis.

By means of illustration, Fig. 1 shows the process whereby proteomes are generated from normal and disease samples and how differentially expressed proteins are identified. The potential of this type of analysis is tremendous. For example, from a mammalian cell sample, in excess of 2000 proteins can typically be resolved within the proteome. The quality of this is shown in Fig. 2, which shows representative proteomes from three diverse biological sources: human serum, the pathogenic fungus *Candida albicans* and the human hepatoma cell line Huh7.

Use of proteomics to identify disease specific proteins

In most cases, the drug discovery process is initiated by the identification of a novel candidate target – almost always a protein – that is believed to be instrumental in the disease process. To date, there is a variety of means whereby drug targets have been forthcoming. These include molecular, cellular and genomic approaches, mostly centred upon DNA and mRNA analysis. The gene in question is isolated, and expression and characterization of its coded protein product – i.e. the drug target – is invariably a secondary event.

With the proteomic approach, the starting point is at the other end of the 'telescope'. Here there is direct and im-

mediate comparison of the proteomes from paired normal and disease materials. Examples of these pairs are: (1) purified epithelial cell populations derived from human breast tumours, matched to purified normal populations of human breast epithelial cells, and (2) the invading pathogenic hyphal form of *C. albicans*, matched to the non-invading yeast form of *C. albicans*. When the proteome images from each pair are aligned, the Proteograph™ software is able to rapidly identify those proteins (each referenced as having a unique molecular cluster index, or MCI) that are either unique, or those that are differentially expressed. Thus, the Proteograph output from this analysis is both qualitative and quantitative.

Proteograph analysis for a particular study can also be undertaken on any number of samples. For example, one might compare anything from a few to several hundred preparations or samples, each from a normal and disease counterpart, and have these analysed in a single Proteograph study. In this way, it is possible to assign strong statistical confidence to the data and in some instances to identify specific subpopulations within the input biological sources. This feature will become increasingly significant in the near future, and there is a clear synergy here whereby proteomics can work closely with pharmacogenomic approaches to stratify patient populations and achieve effective targeted care for the patient. Whatever the source of the materials, the net output of Proteograph analysis is immediate identification of disease specific proteins. This is shown in Fig. 3, which shows the results of a proteograph obtained by comparing untreated human hepatoma cells with cells following exposure to a clinical

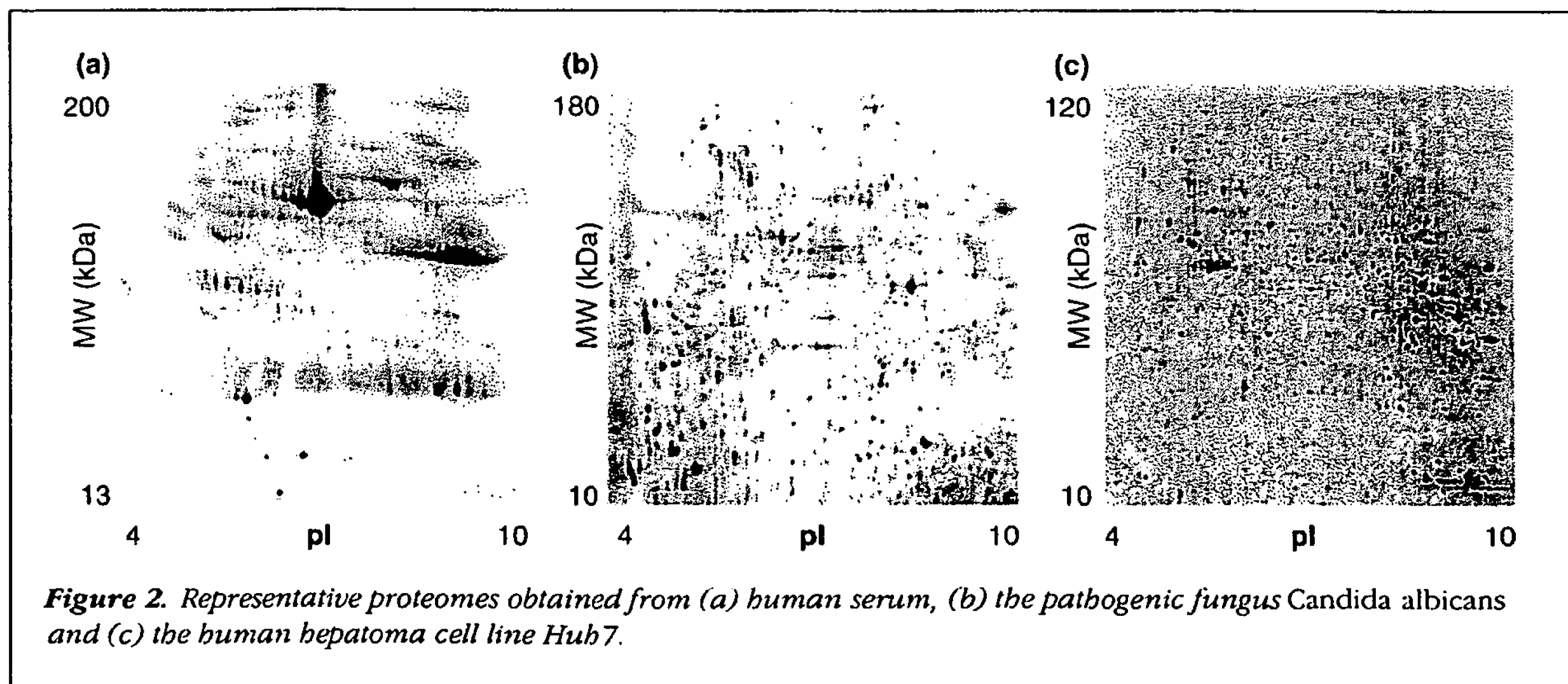
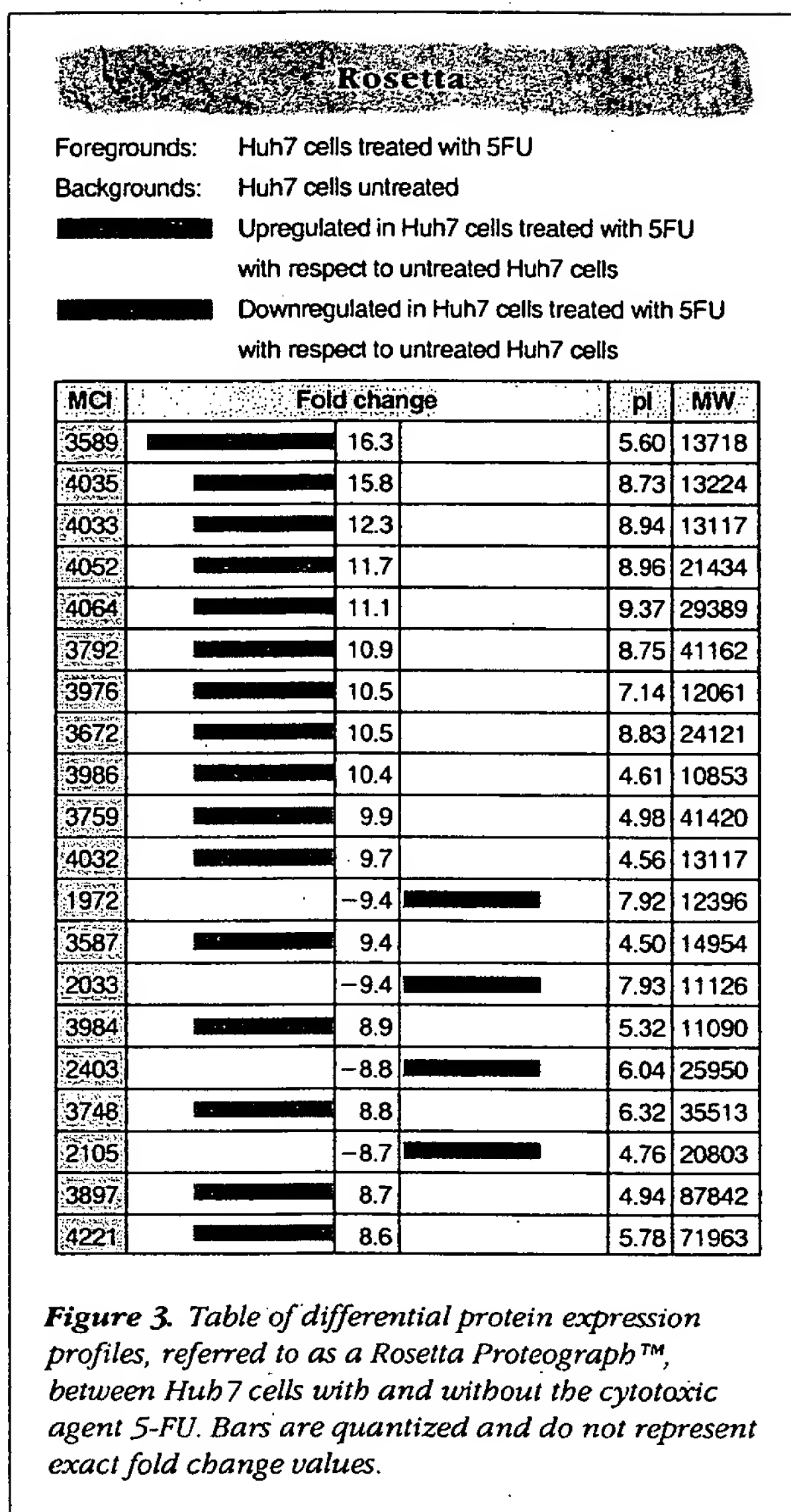


Figure 2. Representative proteomes obtained from (a) human serum, (b) the pathogenic fungus *Candida albicans* and (c) the human hepatoma cell line Huh7.



cytotoxic agent. In this instance, only the top 20 differentially expressed MCIs are shown, but the readout would normally extend to a defined cut-off value, typically a two-fold or greater difference in expression levels, determined by the user.

In a typical analysis involving disease and normal mammalian material, in which each proteome would have ~2000 protein features each assigned an MCI, the proteograph might identify somewhere in the region of 50–300 MCIs that are unique or differentially expressed. To capitalize rapidly on these data, at OGS a high-throughput

mass spectrometry facility coupled to advanced databases to annotate these MCIs as individual proteins is applied. As these are all disease specific proteins, each could represent a novel target and/or a novel disease marker. The process becomes even more powerful when a panel of features, rather than individual features, are assigned. The relevance of this is apparent when one considers that most diseases, if not all, are multifactorial in nature and arise from polygenic changes. Rather than analysing events in isolation, the ability to examine hundreds or thousands of events simultaneously, as shown by proteomics, can offer real advantages.

Identification and assignment of candidate targets

The rapid identification and assignment of candidate targets and markers represents a huge challenge, but this has been greatly facilitated by combining the recent advances made in proteomics and analytical mass spectrometry⁹. Using automated procedures it is now possible to annotate proteins present in femtomole quantities, which would depict the low abundance class of proteins. The process of annotation is similarly aided by the quality and richness of the sequence specific databases that are currently available, both in the public domain and in the private sector (e.g. those supplied by Incyte Pharmaceuticals). In this respect, the advances in proteomics have benefited considerably from the breakthroughs achieved with genomics.

From an application perspective, cancer studies provide a good opportunity whereby proteomics can be instrumental in identifying disease specific proteins, because it is often feasible to obtain normal and diseased tissue from the same patient. For example, proteomic studies have been reported on neuroblastomas¹⁰, human breast proteins from normal and tumour sources^{11–13}, lung tumours¹⁴, colon tumours¹⁵ and bladder tumours¹⁶. There are also proteomic studies reported within the cardiovascular therapeutic area, in which disease or response proteins are identified^{17,18}.

Genomic microarray analysis can similarly identify unique species or clusters of mRNAs that are disease specific. However, in some instances, there is a clear lack of correlation between the levels of a specific mRNA and its corresponding protein (Ref. 19, Gypi, S.P. *et al.*, submitted). This has now been noted by many investigators and reaffirms that post-transcriptional events, including protein stability, protein modification (such as phosphorylation, glycosylation, acylation and methylation) and cell localization, can constitute major regulatory steps. Proteomic analysis captures all of these steps and can therefore provide unique and valuable information independent from, or complementary to, genomic data.

Proteomics for target validation and signal transduction studies

The identification of disease specific proteins alone is insufficient to begin a drug screening process. It is critical to assign function and validation to these proteins by confirming they are indeed pivotal in the disease process. These studies need to encompass both gain- and loss-of-function analyses. This would determine whether the activity of a candidate target (an enzyme, for example), eliminated by molecular/cellular techniques, could reverse a disease phenotype. If this happened, then the investigator would have increased confidence that a small-molecule inhibitor against the target would also have a similar effect. The proposal of candidate drug targets is often not a difficult process, but validating them is another matter. Validation represents a major bottleneck where the wrong decision can have serious consequences²⁰.

Proteomics can be used to evaluate the role of a chosen target protein in signal transduction cascades directly relevant to the disease. In this manner, valuable information is forthcoming on the signalling pathways that are perturbed by a target protein and how they might be corrected by appropriate therapeutics. Techniques that are well established in one-dimensional protein studies to investigate signalling pathways, such as western blotting and immunoprecipitation, are highly suited to proteomic applications. For example, the proteomes obtained can be blotted onto membranes and probed with antibodies against the target protein or related signalling molecules²¹⁻²³. Because proteomics can resolve >2000 proteins on a single gel, it is possible to derive important information on specific isoforms (such as glycosylated or phosphorylated variants) of signalling molecules. This will result in characterization of how they are altered in the disease process. Western immunoblotting techniques using high-affinity antibodies will typically identify proteins present at ~10 copies per cell (~1.7 fmol); this is in contrast to the best fluorescent dyes currently available that are limited to imaging proteins at 1000 or more copies per cell. The level of sensitivity derived by these applications will greatly facilitate interpretation of complex signalling pathways and contribute significantly to validation of the target under study.

Immunoprecipitation studies

Similarly, immunoprecipitation studies are another useful way to exploit the resolving power of proteomics^{24,25}. In this instance, very large quantities of protein (e.g. several milligrams) can be subjected to incubation with antibodies against chosen signalling molecules. This allows high-affin-

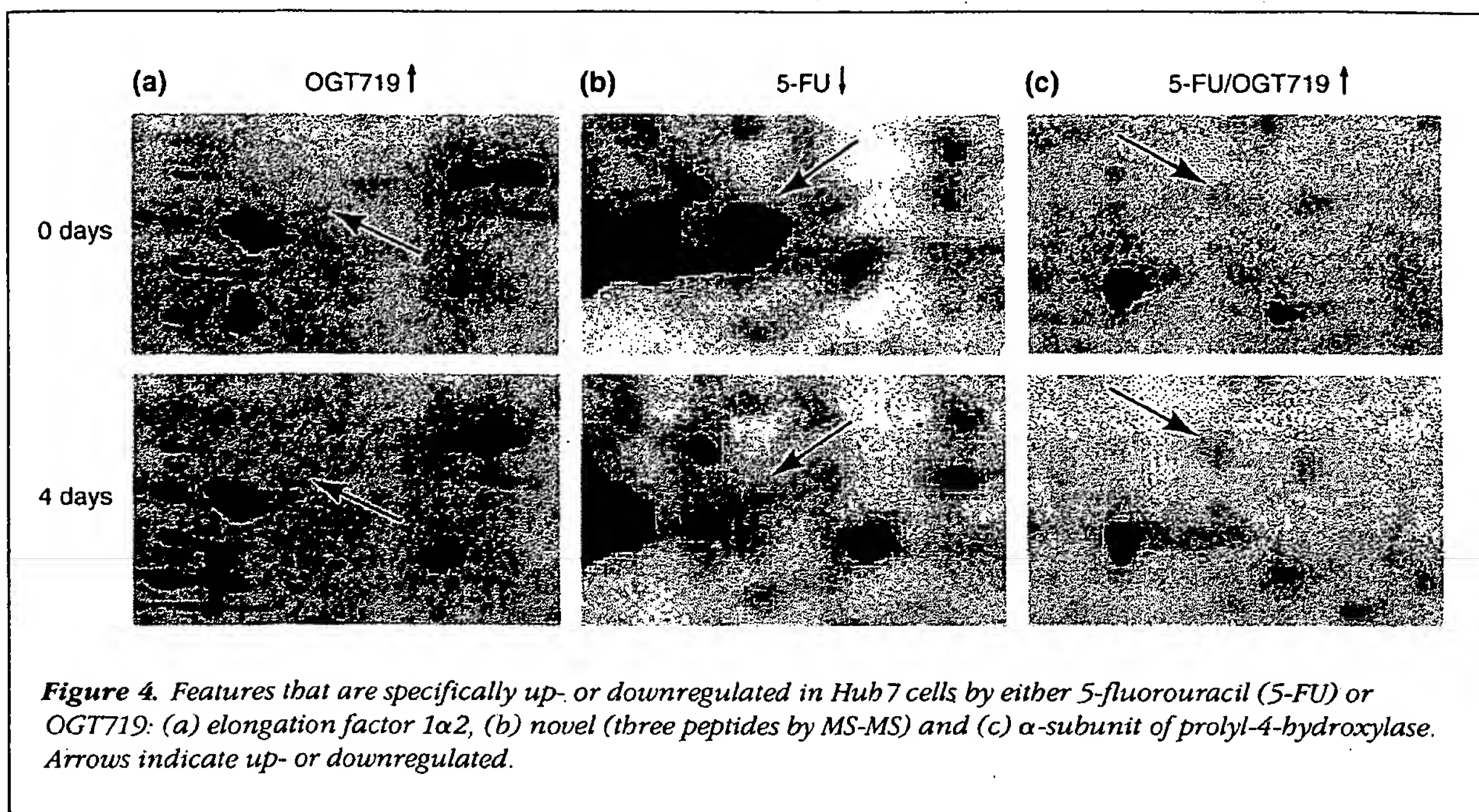
ity capture of these proteins, which can subsequently be eluted and electrophoresed on a 2D gel to provide a high-resolution proteome of a specific subset of proteins. Detection by blot analysis allows the identification of extremely small amounts of defined signalling molecules. Again, the different isoforms of even very low abundance proteins can be seen, and, very importantly, the technique allows the investigator to identify multiprotein complexes or other proteins that co-precipitate with the target protein. These coassociating proteins frequently represent signalling partners for the target protein, and their identification by mass spectrometry can lead to invaluable information on the signalling processes involved.

The depth of signal transduction analysis offered by proteomics, and the utility for target validation studies, can be extended even further by applying cell fractionation studies²⁶⁻²⁸. By purifying subcellular fractions, such as membrane, nuclear, organelle and cytosolic, it is possible to assign a localization to proteins of interest and to follow their trafficking in a cell. Enrichment of these fractions will also allow much higher representation of low abundance proteins on the proteome. Their detection by fluorescent dyes or immunoblot techniques will lead to the identification of proteins in the range of 1-10 copies per cell, putting the sensitivity on a par with genomic approaches.

These signal transduction analyses can be of additional value in experiments where inhibitors derived from a screening programme against the target are being evaluated for their potency and selectivity. The inhibitors can encompass small molecules, antisense nucleic acid constructs, dominant-negative proteins, or neutralizing antibodies microinjected into cells. In each case, proteome analysis can provide unique data in support of validation studies for a chosen candidate drug target.

Proteomics and drug mode-of-action studies

Once a validated target is committed to a screening regimen to identify and advance a lead molecule, it is important to confirm that the efficacy of the inhibitor is through the expected mechanism. Such mode-of-action studies are usually tackled by various cell biological and biochemical methods. Proteomics can also be usefully applied to these studies and this is illustrated below by describing data obtained with OGT719. This is a novel galactosyl derivative of the cytotoxic agent 5-fluorouracil (5-FU), which is currently being developed by OGS for the treatment of hepatocellular carcinoma and colorectal metastases localized in the liver. The premise underpinning the design and rationale of OGT719 was to derive a 5-FU prodrug capable



of targeting, and being retained in, cells bearing the asialoglycoprotein receptor (ASGP-r), including hepatocytes²⁹, hepatoma Huh7 cells³⁰ and some colorectal tumour cells³¹. The growth of the human hepatoma cell line Huh7 is inhibited by 5-FU or by OGT719. If the inhibition by OGT719 were the result of uptake and conversion to 5-FU as the active component, then it would be expected that Huh7 cells would show similar proteome profiles following exposure to either drug.

To examine these possibilities, we conducted an experiment taking samples of Huh7 cells that had been treated with IC_{50} doses of either OGT719 or 5-FU. Total cell lysates were prepared and taken through 2D electrophoresis, fluorescence staining, digital imaging and Proteograph analysis. To facilitate the interpretation of the data across all of the 2291 features seen on the proteomes, drug-induced protein changes of fivefold or greater, identified by the Proteograph, were analysed further. Interestingly, from this analysis 19 identical proteins were changed fivefold or more by both drugs, strongly suggesting similarities in the mode of action for these two compounds.

Thus, from very complex data involving >2000 protein features, using proteomics it is possible to analyse quantitatively and qualitatively each protein during its exposure to drugs. The biologist is now able to focus a series of further studies specifically on an enriched subset of proteins.

Figure 4 shows highlighted examples of the selected areas of the proteome where some of these identified proteins in the above study are altered in response to either or both drugs.

Several of the proteins identified above as being modulated similarly by 5-FU or OGT719 in Huh7 cells were subjected to tandem mass-spectrometric analysis for annotation. Some of these, such as the nuclear ribosomal RNA-binding protein³², can be placed into pyrimidine pathways or related cell cycle/growth biochemical pathways in which 5-FU is known to act.

To attribute further significance to the proteome mode-of-action studies with OGT719, another cell line, the rat sarcoma HSN, was used. Growth of these cells is inhibited by 5-FU, but they are completely refractory to OGT719; notably they lack the ASGP-r, which might explain this finding (unpublished). For our proteome studies, HSN cells were treated with 5-FU or OGT719 over a time course of one, two and four days. At each time point, cells were harvested and processed to derive proteomes and Proteographs. As before, we purposely focused on those proteins that increased or decreased by fivefold or more. In this instance, there were no proteins co-modulated by the two drugs. This is perhaps to be expected, given that the HSN cells are killed by 5-FU and yet are refractory to OGT719.

Clear potential

The above is just an example of how proteomics can be used to address the mode of action of anticancer drugs. The potential of this approach is clear, and one can envisage situations where it will be profitable to compare the proteomes of cells in which the drug target has been eliminated by molecular knockout techniques, or with small-molecule inhibitors believed to act specifically on the same target. In addition to using proteomics to examine the action of drugs, it is also possible to use this approach to gauge the extent of nonspecific effects that might eventually lead to toxicity. For instance, in the example used above with HSN cells treated with OGT719, although cell growth was not affected, the levels of several specific proteins were changed. Further investigation of these proteins and the signalling pathways in which they are involved could be illuminating in predicting the likelihood or otherwise of long-term toxicity.

Use of proteomics in formal drug toxicology studies

A drug discovery programme at the stage where leads have been identified and mode-of-action studies are advanced, will proceed to investigate the pharmacokinetic and toxicology profile of those agents. These two parameters are of major importance in the drug discovery process, and many agents that have looked highly promising from *in vitro* studies have subsequently failed because of insurmountable pharmacokinetic and/or toxicity problems *in vivo*. Whereas the pharmacokinetic properties of a molecule can now be characterized quickly and accurately, toxicity studies are typically much longer and more demanding in their interpretation.

The ability to achieve fast and accurate predictions of toxicity within an *in vivo* setting would represent a big step forward in accelerating any drug discovery programme. Toxicity from a drug can be manifested in any organ. However, because the liver and kidney are the major sites in the body responsible for metabolism and elimination of most drugs, it is informative to examine these particular organs in detail to provide early indications about events that might result in toxicity.

The basis for most xenobiotic metabolizing activity is to increase the hydrophilicity of the compound and so facilitate its removal from the body. Most drugs are metabolized in the liver via the cytochrome P450 family of enzymes, which are known to comprise a total of ~200 different members^{33,34}, encompassing a wide array of overlapping specificities for different substrates. In addition to clearance, they also play a major role in metabo-

lism that can lead to the production and removal of toxic species, and in some instances it is possible to correlate the ability or failure to remove such a toxin with a specific P450 or subgroup.

Unique P450 profiles

Each individual person will have a slightly different P450 profile, largely from polymorphisms and changes in expression levels, although other genetic and environmental factors aside from P450 also need to be taken into consideration. A significant amount of research is currently being directed towards this field – known as pharmacogenomics – with the aim of predicting how a patient will respond to a drug, as determined by their genetic make-up^{35–37}. The marked variation of individuals in their ability to clear a compound can be one of the key factors in deciding the overall pharmacokinetic profile of a drug. Not only will this have a bearing on the likelihood of a patient responding to a treatment, but it will also be a factor in determining the possibility of their experiencing an adverse effect.

Many pharmaceutical companies are already employing genomic approaches, involving P450 measurements, as a key step in their assessment of the toxicological profile of a candidate drug and therefore of its suitability, or otherwise, to be considered for human clinical trials. There are limits to this approach, however. Whereas the P450 mRNA profiling can predict with some accuracy the likely metabolic fate of a drug, it will not provide information on whether the metabolites would subsequently lead to toxicity. Besides the patient-to-patient differences in steady-state levels of the P450s, there are also characteristic induction responses of these enzymes to some drugs. Moreover, as there can be some doubt over the correlation of mRNA levels and the corresponding protein levels, there is scope for misinterpretation of the results and hence real advantages to be gained from a proteome approach. In both instances, the ability to examine entire proteome profiles, including the P450 proteins, will be a significant advantage in understanding and predicting the metabolism and toxicological outcome of drugs.

In addition to direct organ and tissue studies, the serum, which collects the majority of toxicity markers released from susceptible organs and tissues throughout the entire body, can be utilized. Serum is rich in nuclease activity and, as pharmacogenomics is not suited to deal with these samples, valuable markers of toxicity could go undetected. However, by using proteomics for these types of analyses, serum markers (and clusters thereof) are now accessible for evaluation as indicators of toxicity.

Pharmacoproteomics

Proteomics can thus be used to add a new sphere of analysis to the study of toxicity at the protein level, and in the era of '-omics' there is a case to be made to adopt the term 'Pharmacoproteomics™'. Animals can be dosed with increasing levels of an experimental drug over time, and serum samples can be drawn for consecutive proteome analyses. Using this procedure, it should be possible to identify individual markers, or clusters thereof, that are dose related and correlate with the emergence and severity of toxicity. Markers might appear in the serum at a defined drug dose and time that are predictive of early toxicity within certain organs and if allowed to continue will have damaging consequences. These serum markers could subsequently be used to predict the response of each individual and allow tailoring of therapy whereby optimal efficacy is achieved without adverse side effects being apparent. This application can obviously extend to tracking toxicity of drugs in clinical trials where serum can be readily drawn and analysed. Surrogate markers for drug efficacy could also be detected by this procedure and could facilitate the challenge of identifying patient classes who will respond favourably to a drug and at what dosage.

Conclusions

By contrast to the agents administered to patients in clinical wards, the process of drug discovery is not a prescriptive series of steps. The risks are high and there are long timelines to be endured before it is known whether a candidate drug will succeed or fail. At each step of the drug discovery process there is often scope for flexibility in interpretation, which over many steps is cumulative. The pharmaceutical companies most likely to succeed in this environment are those that are able to make informed accurate decisions within an accelerated process.

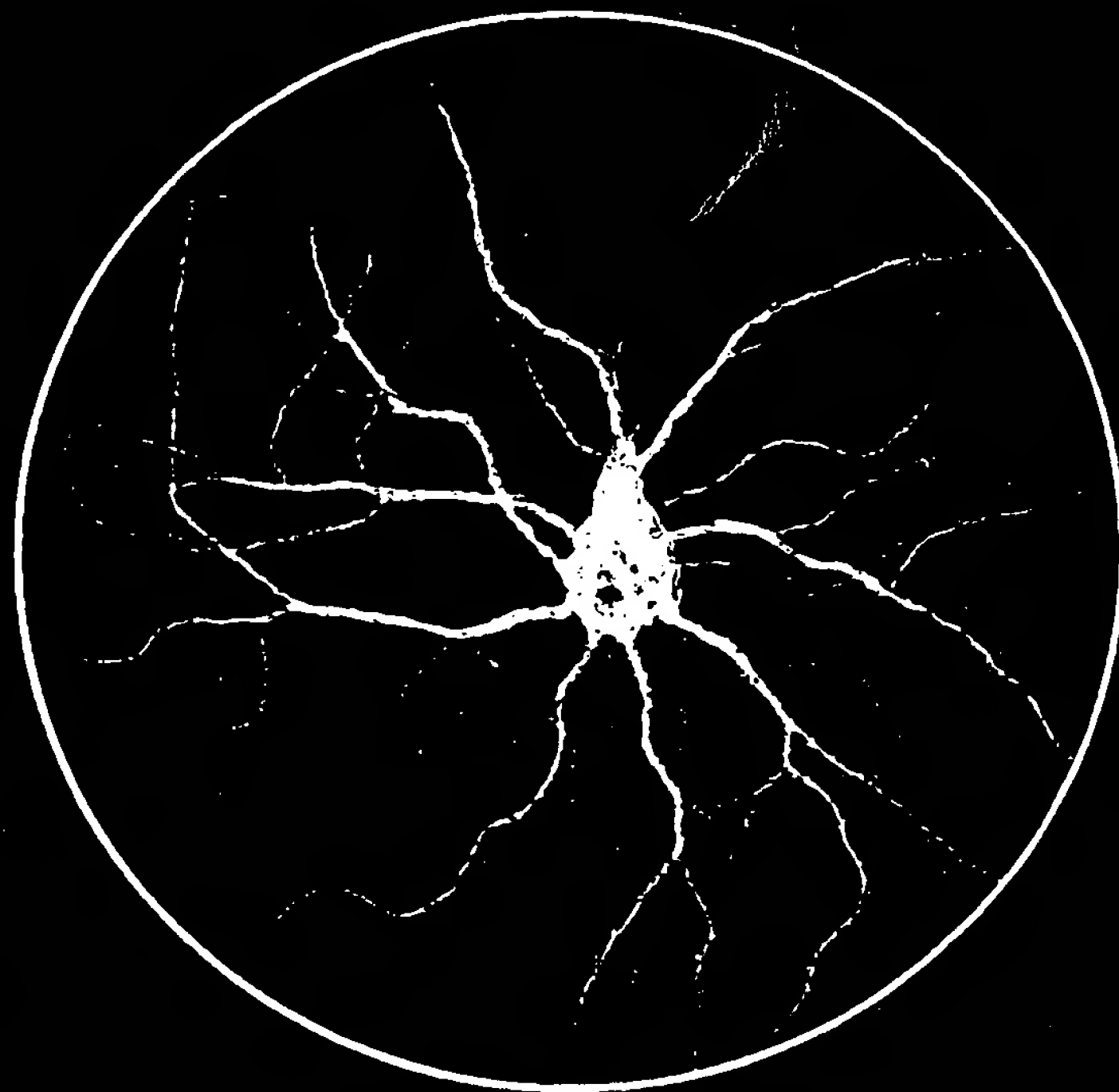
The genomics revolution has impacted very positively upon these issues and now has a powerful new partner in proteomics. The ability to undertake global analysis of proteins from a very wide diversity of biological systems and to interrogate these in a high-throughput, systematic manner will add a significant new dimension to drug discovery. Each step of the process from target discovery to clinical trials is accessible to proteomics, often providing unique sets of data. Using the combination of genomics and proteomics, scientists can now see every dimension of their biological focus, from genes, mRNA, proteins and their subcellular localization. This will greatly assist our understanding of the fundamental mechanistic basis of human disease and allow new improved and speedier drug discovery strategies to be implemented.

REFERENCES

- 1 Crooke, S.T. (1998) *Nat. Biotechnol.* 16, 29–30
- 2 Dykes, C.W. (1996) *Br. J. Clin. Pharmacol.* 42, 683–695
- 3 Schena, M. *et al.* (1998) *Trends Biotechnol.* 16, 301–306
- 4 Ramsay, G. (1998) *Nat. Biotechnol.* 16, 40–44
- 5 Anderson, N.L. and Anderson, N.G. (1998) *Electrophoresis* 19, 1853–1861
- 6 James, P. (1997) *Biochem. Biophys. Res. Commun.* 231, 1–6
- 7 Wilkins, M.R. *et al.* (1996) *Biotechnol. Genet. Eng. Rev.* 13, 19–50
- 8 Parekh, R.B. and Rohlf, C. (1997) *Curr. Opin. Biotechnol.* 8, 718–723
- 9 Figey, D. *et al.* (1998) *Electrophoresis* 19, 1811–1818
- 10 Wimmer, K. *et al.* (1996) *Electrophoresis* 17, 1741–1751
- 11 Giometti, C.S., Williams, K. and Tollaksen, S.L. (1997) *Electrophoresis* 18, 573–581
- 12 Williams, K. *et al.* (1998) *Electrophoresis* 19, 333–343
- 13 Rasmussen, R.K. *et al.* (1998) *Electrophoresis* 19, 818–825
- 14 Hirano, T. *et al.* (1995) *Br. J. Cancer* 72, 840–848
- 15 Ji, H. *et al.* (1997) *Electrophoresis* 18, 605–613
- 16 Ostergaard, M. *et al.* (1997) *Cancer Res.* 57, 4111–4117
- 17 Patel, V.B. *et al.* (1997) *Electrophoresis* 18, 2788–2794
- 18 Arnott, D. *et al.* (1998) *Anal. Biochem.* 258, 1–18
- 19 Anderson, L. and Seilhamer, J. (1997) *Electrophoresis* 18, 533–537
- 20 Rastan, S. and Beeley, L.J. (1997) *Curr. Opin. Genet. Dev.* 7, 777–783
- 21 Gravel, P. *et al.* (1995) *Electrophoresis* 16, 1152–1159
- 22 Qian, Y. *et al.* (1997) *Clin. Chem.* 43, 352–359
- 23 Sanchez, J.C. *et al.* (1997) *Electrophoresis* 18, 638–641
- 24 Watts, A.D. *et al.* (1997) *Electrophoresis* 18, 1086–1091
- 25 Asker, N. *et al.* (1995) *Biochem. J.* 308, 873–880
- 26 Ramsby, M.L., Makowski, G.S. and Khairallah, E.A. (1994) *Electrophoresis* 15, 265–277
- 27 Huber, L.A. (1995) *FEBS Lett.* 369, 122–125
- 28 Corthals, G.L. *et al.* (1997) *Electrophoresis* 18, 317–323
- 29 Hubbard, A.L., Wall, D.A. and Ma, A. (1983) *J. Cell Biol.* 96, 217–229
- 30 Zeng, F.Y., Oka, J.A. and Weigel, P.H. (1996) *Biochem. Biophys. Res. Commun.* 218, 325–330
- 31 Mu, J.-Z. *et al.* (1994) *Biochim. Biophys. Acta* 1222, 483–491
- 32 Ghoshal, K. and Jacob, S.T. (1997) *Biochem. Pharmacol.* 53, 1569–1575
- 33 Guengerich, F.P. and Parikh, A. (1997) *Curr. Opin. Biotechnol.* 8, 623–628
- 34 Rendic, S. and Di Carlo, F.J. (1997) *Drug Metab. Rev.* 29, 413–580
- 35 Vermees, A., Guchelaar, H.J. and Koopmans, R.P. (1997) *Cancer Treat. Rev.* 23, 321–339
- 36 Housman, D. and Ledley, F.D. (1998) *Nat. Biotechnol.* 16, 492–493
- 37 Persidis, A. (1998) *Nat. Biotechnol.* 16, 209–210

MOLECULAR BIOLOGY OF THE CELL

THIRD EDITION



Bruce Alberts • Dennis Bray
Julian Lewis • Martin Raff
Keith Roberts • James D. Watson



extracts. If these minor cell proteins differ among cells to the same extent as the more abundant proteins, as is commonly assumed, only a small number of protein differences (perhaps several hundred) suffice to create very large differences in cell morphology and behavior.

A Cell Can Change the Expression of Its Genes in Response to External Signals³

Most of the specialized cells in a multicellular organism are capable of altering their patterns of gene expression in response to extracellular cues. If a liver cell is exposed to a glucocorticoid hormone, for example, the production of several specific proteins is dramatically increased. Glucocorticoids are released during periods of starvation or intense exercise and signal the liver to increase the production of glucose from amino acids and other small molecules; the set of proteins whose production is induced includes enzymes such as tyrosine aminotransferase, which helps to convert tyrosine to glucose. When the hormone is no longer present, the production of these proteins drops to its normal level.

Other cell types respond to glucocorticoids in different ways. In fat cells, for example, the production of tyrosine aminotransferase is reduced, while some other cell types do not respond to glucocorticoids at all. These examples illustrate a general feature of cell specialization—different cell types often respond in different ways to the same extracellular signal. Underlying this specialization are features that do not change, which give each cell type its permanently distinctive character. These features reflect the persistent expression of different sets of genes.

Gene Expression Can Be Regulated at Many of the Steps in the Pathway from DNA to RNA to Protein⁴

If differences between the various cell types of an organism depend on the particular genes that the cells express, at what level is the control of gene expression exercised? There are many steps in the pathway leading from DNA to protein, and all of them can in principle be regulated. Thus a cell can control the proteins it makes by (1) controlling when and how often a given gene is transcribed (**transcriptional control**), (2) controlling how the primary RNA transcript is spliced or otherwise processed (**RNA processing control**), (3) selecting which completed mRNAs in the cell nucleus are exported to the cytoplasm (**RNA transport control**), (4) selecting which mRNAs in the cytoplasm are translated by ribosomes (**translational control**), (5) selectively destabilizing certain mRNA molecules in the cytoplasm (**mRNA degradation control**), or (6) selectively activating, inactivating, or compartmentalizing specific protein molecules after they have been made (**protein activity control**) (Figure 9-2).

For most genes transcriptional controls are paramount. This makes sense because, of all the possible control points illustrated in Figure 9-2, only transcriptional control ensures that no superfluous intermediates are synthesized. In the

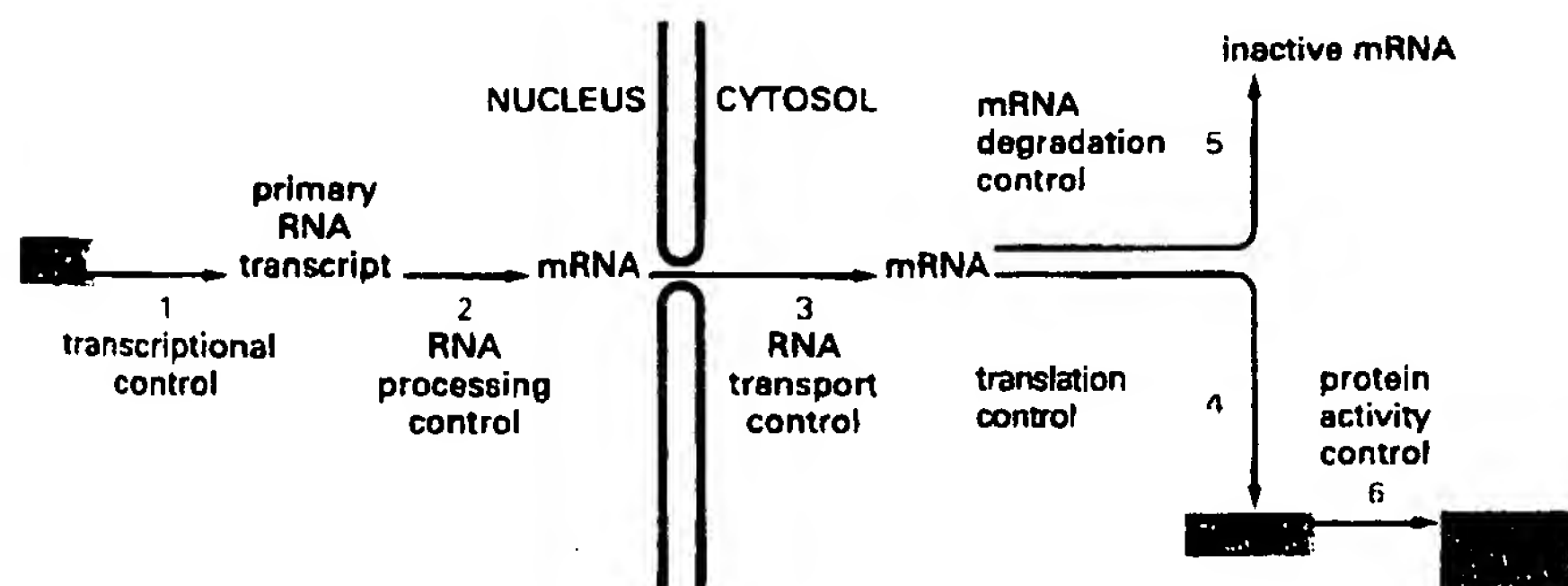


Figure 9-2 Six steps at which eucaryote gene expression can be controlled. Only controls that operate at steps 1 through 5 are discussed in this chapter. The regulation of protein activity (step 6) is discussed in Chapter 5; this includes reversible activation or inactivation by protein phosphorylation as well as irreversible inactivation by proteolytic degradation.

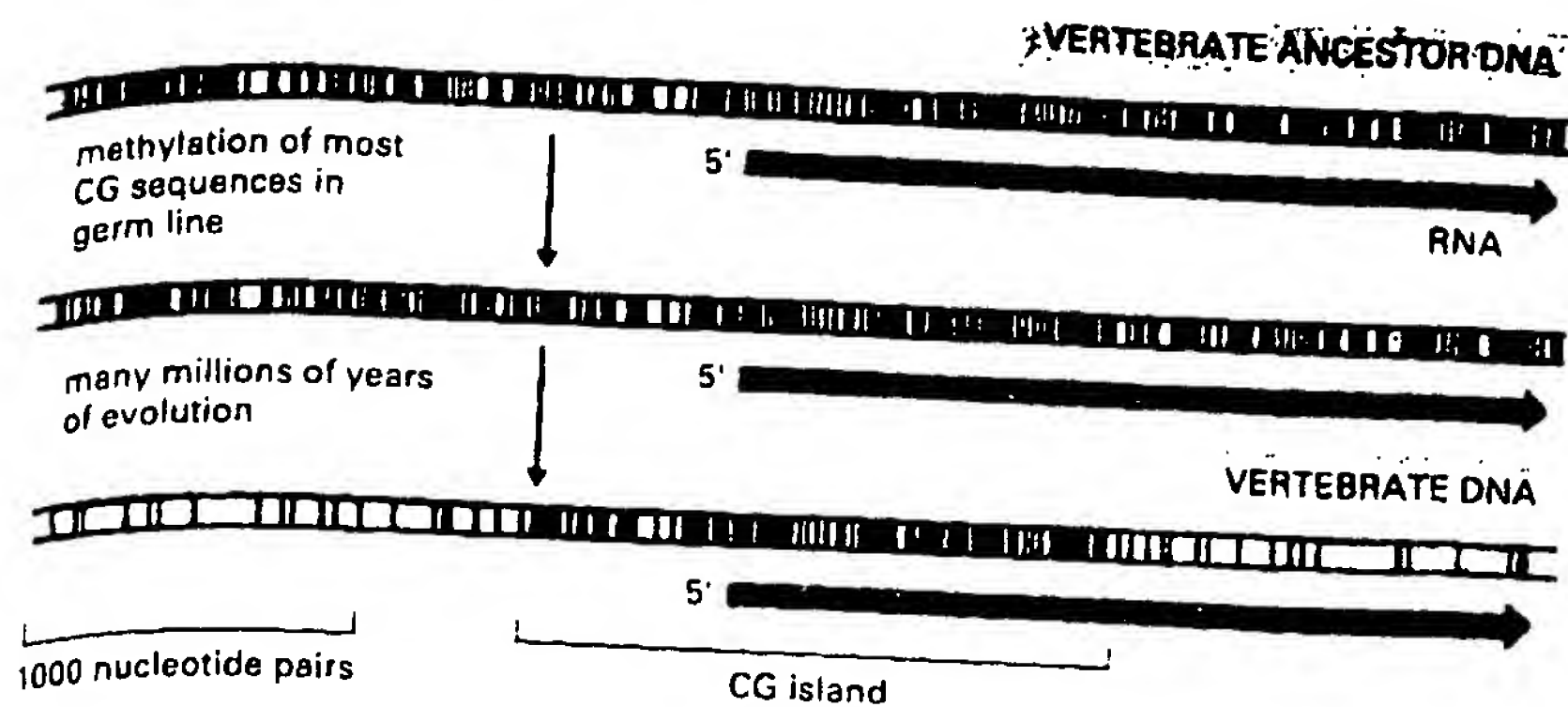


Figure 9-71 A mechanism to explain both the marked deficiency of CG sequences and the presence of CG islands in vertebrate genomes. A black line marks the location of an unmethylated CG dinucleotide in the DNA sequence, while a red line marks the location of a methylated CG dinucleotide.

Summary

The many types of cells in animals and plants are created largely through mechanisms that cause different genes to be transcribed in different cells. Since many specialized animal cells can maintain their unique character when grown in culture, the gene regulatory mechanisms involved in creating them must be stable once established and heritable when the cell divides, endowing the cell with a memory of its developmental history. Prokaryotes and yeasts provide unusually accessible model systems in which to study gene regulatory mechanisms, some of which may be relevant to the creation of specialized cell types in higher eucaryotes. One such mechanism involves a competitive interaction between two (or more) gene regulatory proteins, each of which inhibits the synthesis of the other; this can create a flip-flop switch that switches a cell between two alternative patterns of gene expression. Direct or indirect positive feedback loops, which enable gene regulatory proteins to perpetuate their own synthesis, provide a general mechanism for cell memory.

In eucaryotes gene transcription is generally controlled by combinations of gene regulatory proteins. It is thought that each type of cell in a higher eucaryotic organism contains a specific combination of gene regulatory proteins that ensures the expression of only those genes appropriate to that type of cell. A given gene regulatory protein may be expressed in a variety of circumstances and typically is involved in the regulation of many genes.

In addition to diffusible gene regulatory proteins, inherited states of chromatin condensation are also utilized by eucaryotic cells to regulate gene expression. In vertebrates DNA methylation also plays a part, mainly as a device to reinforce decisions about gene expression that are made initially by other mechanisms.

Posttranscriptional Controls

Although controls on the initiation of gene transcription are the predominant form of regulation for most genes, other controls can act later in the pathway from RNA to protein to modulate the amount of gene product that is made. Although these **posttranscriptional controls**, which operate after RNA polymerase has bound to the gene's promoter and begun RNA synthesis, are less common than *transcriptional control*, for many genes they are crucial. It seems that every step in gene expression that could be controlled in principle is likely to be regulated under some circumstances for some genes.

We consider the varieties of posttranscriptional regulation in temporal order, according to the sequence of events that might be experienced by an RNA molecule after its transcription has begun (Figure 9-72).

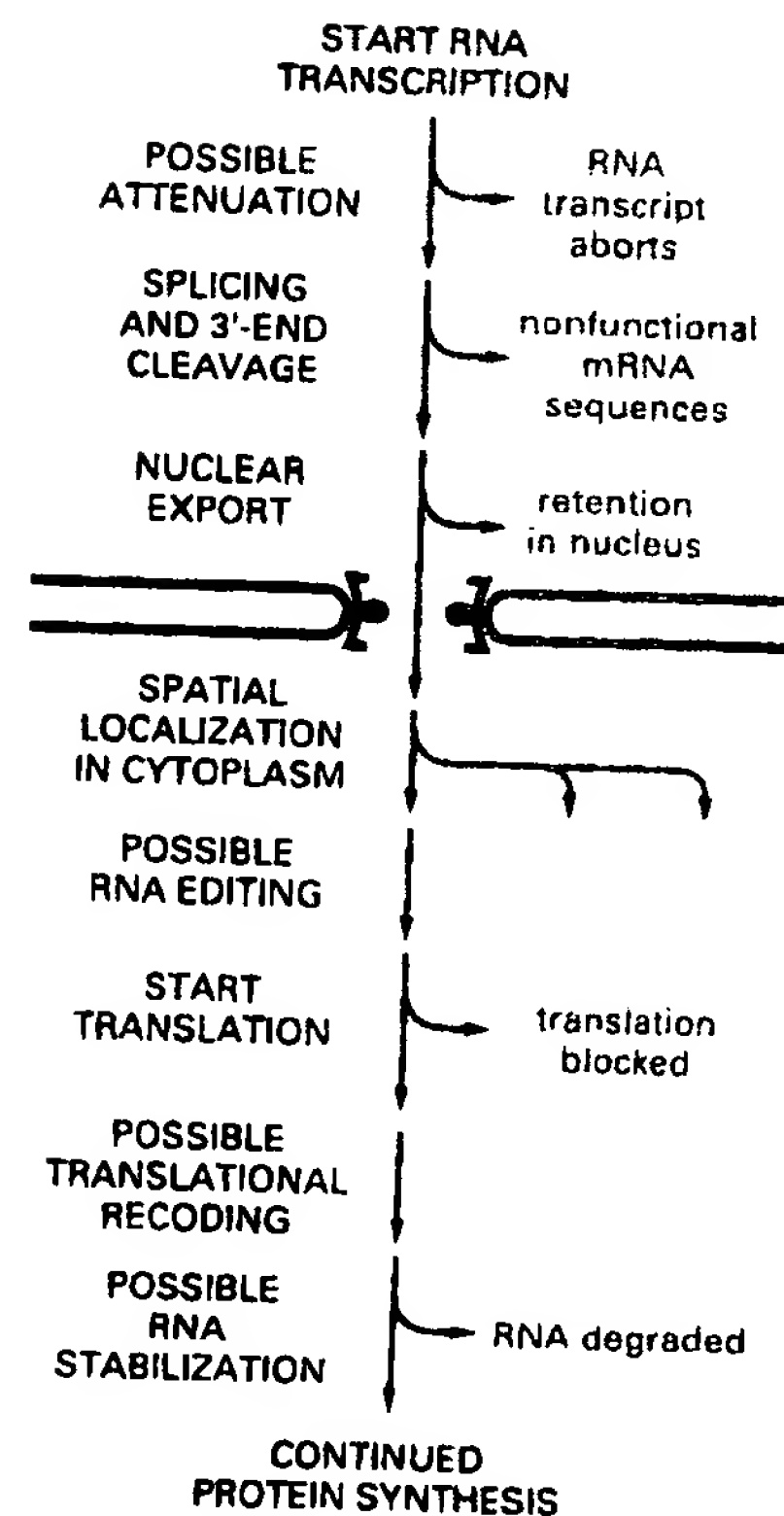
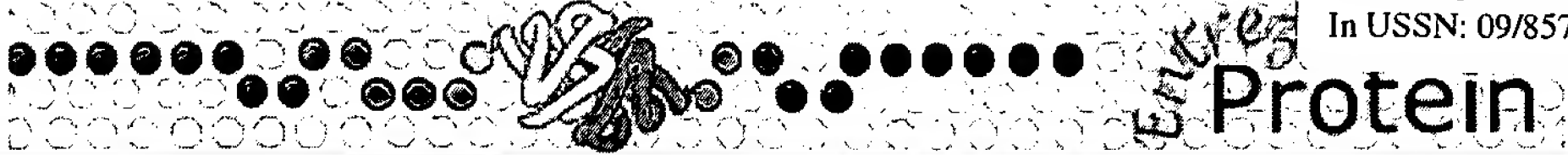


Figure 9-72 Possible post-transcriptional controls on gene expression. Only a few of these controls are likely to be used for any one gene.



EntrezPubMedNucleotideProteinGenomeStructurePMCTaxonomyBooks

SearchProtein▼forGoClear

LimitsPreview/IndexHistoryClipboardDetails

Displaydefault▼Show:20▼Send toFile▼Get SubsequenceFe

☐ 1: AAC52025. clone 22 [Homo sa...[gi:2271473]

BLink, Links

LOCUS AAC52025 248 aa linear PRI 17-FEB-1998

DEFINITION clone 22 [Homo sapiens].

ACCESSION AAC52025

VERSION AAC52025.1 GI:2271473

DBSOURCE locus AF009426 accession AF009426.1

KEYWORDS .

SOURCE Homo sapiens (human)

ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.

REFERENCE 1 (residues 1 to 248)

AUTHORS Yoshikawa,T., Sanders,A.R., Esterling,L.E., Overhauser,J., Garnes,J.A., Lennon,G., Grewal,R. and Detera-Wadleigh,S.D.

TITLE Isolation of chromosome 18-specific brain transcripts as positional candidates for bipolar disorder

JOURNAL Am. J. Med. Genet. 74 (2), 140-149 (1997)

MEDLINE 97275951

PUBMED 9129712

REFERENCE 2 (residues 1 to 248)

AUTHORS Yoshikawa,T., Sanders,A.R., Esterling,L.E. and Detera-Wadleigh,S.D.

TITLE Multiple transcriptional variants and RNA editing in C18orf1, a novel gene with LDLRA and transmembrane domains on 18p11.2

JOURNAL Genomics 47 (2), 246-257 (1998)

MEDLINE 98140124

PUBMED 9479497

REFERENCE 3 (residues 1 to 248)

AUTHORS Yoshikawa,T. and Detera-Wadleigh,S.D.

TITLE Direct Submission

JOURNAL Submitted (20-JUN-1997) Clinical Neurogenetics Branch, National Institute of Mental Health, Bethesda, MD 20892, USA

COMMENT Method: conceptual translation supplied by author.

FEATURES Location/Qualifiers

source 1..248
/organism="Homo sapiens"
/db_xref="taxon:9606"
/chromosome="18"
/map="18p11.2"

Protein 1..248
/product="clone 22"

CDS 1..248
/coded_by="AF009426.1:243..989"
/note="alternatively spliced; beta-1 form; possible membrane-spanning protein"

ORIGIN

1 maaelefaqi iivvvvtvm vvivcllnh ykvstrsfin rpnqsrred glpgegclwp
61 sdsaaprlga seimhaprsr drftapsfiq rdrfsrfqpt ypyvqheidl pptislsdge
121 epppyqgpct lqlrdpeqqm elnresvrp pnrtifdsdl idiamysggp cppssnsgis
181 astcssngm egppptysev mghhpgasfl hhqrsnahrg srlqfqgna estivpikgk
241 drkpgnlv

//

Confidential -- Property of Incyte Corporation LifeSeq Gold 5.1 Nov 2002

Program: blastp
Sequence ID(s):
1871288CD1 (LGflJAN2002p) vs. genpept138

NCBI-BLASTP 2.2.3 [May-13-2002]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs", Nucleic Acids Res. 25:3389-3402.

Query= 1871288CD1
(252 letters)

Database: genpept138
1,601,536 sequences; 494,245,048 total letters

Searching.....done

Sequences producing significant alignments:	Score (bits)	E Value
---------------------------------------------	-----------------	------------

g2271473 clone 22 [Homo sapiens]	327	1e-88
----------------------------------	-----	-------

>g2271473 clone 22 [Homo sapiens]
Length = 248

Score = 327 bits (838), Expect = 1e-88
Identities = 169/250 (67%), Positives = 190/250 (75%), Gaps = 7/250 (2%)

Query: 2 AELEFVQIIIIIVVMMVMVVITCLLSHYKLSARSFISRHSQGRRREDALSSEGCLWPSE 61
AELEF QIIIIIVVV+ VMVVVI CLL+HYK+S RSFI+R +Q RRRED L EGCLWPS+
Sbjct: 3 AELEFAQIIIIIVVVVTVMVVIVCLLNHYKVSTRSFINRPNQSRREDGLPQEGCLWPSD 62

Query: 62 STVSGNGIPEPQVYAPPRPTDRLAVPPFAQRERFHRFQPTYPYLQHEIDLPTISLSDGE 121
S G E + PR DR P F QR+RF RFQPTYPY+QHEIDLPTISLSDGE
Sbjct: 63 SAAPRLGASE--IMHAPRSRDRFTAPSFIRQDRFSRFQPTYPYVQHEIDLPTISLSDGE 120

Query: 122 EPPPYQGPCTLQLRDPEQQLELNRESVRAPPNRTIFDSDLMDSARL-GGPCPPSSNSGIS 180
EPPPYQGPCTLQLRDPEQQ+ELNRESVRAPPNRTIFDSDL+D A GGPCPPSSNSGIS
Sbjct: 121 EPPPYQGPCTLQLRDPEQQMELNRESVRAPPNRTIFDSDLIDIAMYSGGPCPPSSNSGIS 180

Query: 181 ATCYGSGGRMEGPPPTYSEVIGHYPGSSFQHQSSGPPSLLEGTRLHHTHIAPLESAAIW 240
A+ S GRMEGPPPTYSEV+GH+PG+SF H Q S + G+RL ES +
Sbjct: 181 ASTCSSNGRMEGPPPTYSEVMGHHPGASFLHHQRS---NAHRGSRLQFQQ-NNAESTIVP 236

Query: 241 SKEKDKQKGH 250
K KD++ G+
Sbjct: 237 IKGKDRKPGN 246